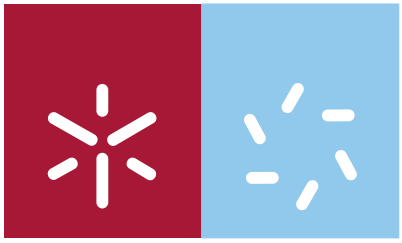
Universidade do Minho
Escola de Ciências

Susana Rafaela Guimarães Martins

Estudo de fatores facilitadores do consumo de publicidade através de *email marketing*

Outubro de 2011

Estudo de fatores facilitadores do consumo de publicidade através de *email marketing*
Susana Rafaela Guimarães Martins
UMinho | 2011



Universidade do Minho
Escola de Ciências

Susana Rafaela Guimarães Martins

Estudo de fatores facilitadores do consumo de publicidade através de *email marketing*

Dissertação de Mestrado
Mestrado em Estatística de Sistemas-Perfil Engenharia e Estatística

Trabalho realizado sob a orientação da
Doutora Cecília Azevedo
e do
Doutor Lino Costa

Outubro de 2011

Nome: Susana Rafaela Guimarães Martins

Endereço electrónico: maleafar@gmail.com Telefone: 964287338

Número do Bilhete de Identidade:12776088

Título de Relatório de Estágio: "Estudo de fatores facilitadores do consumo de publicidade através de *email marketing* "

Orientador(es): Doutora Cecília Azevedo e Doutor Lino Costa

Ano de conclusão: 2011

Designação do Mestrado: Estatística de Sistemas-Perfil Engenharia e Estatística:

Universidade do Minho, ___/___/_____

Assinatura: _____

Resumo

Estudo de fatores facilitadores do consumo de publicidade através de *email marketing*

Este relatório resulta de um projeto desenvolvido numa Empresa muito jovem de Web Marketing sediada no Porto mas que se encontra representada um pouco por todo o mundo.

Foi-nos proposto identificar, através de ferramentas da Estatística, os factores facilitadores do consumo de email marketing.

Depressa nos apercebemos da ambição de tal projeto. Depressa nos apercebemos que não seríamos capazes de dar **a** resposta a este problema. Tentamos aproximações às soluções do problema.

As soluções, apresentadas neste trabalho de síntese, passam pelo entendimento do problema no contexto em que este se insere, marketing/email marketing, pela seleção das variáveis disponíveis na base de dados da Empresa e pela sua análise.

Aplicamos técnicas descritivas, univariadas e multivariadas e técnicas inferenciais.

Usamos técnicas paramétricas, muito poucas, mas tivemos essencialmente que usar técnicas não paramétricas. Estas são apresentadas e justificadas no presente relatório e, naturalmente, são aplicadas ao estudo desenvolvido.

Palavras chave: email marketing, teste de qui-quadrado, normalidade, teste de Kruskal-Wallis, comparações múltiplas, análise de correspondências.

Abstract

Study of factors that facilitate the consumption of advertising through *email marketing*

This report results from a project developed in a very young Web Marketing Company based in Oporto, but it is represented all over the world.

We were proposed to identify, through statistical tools, the factors that facilitate the consumption of email marketing.

We soon came to realize the ambition of such project. Quickly realize that we were not able to give the answer to this problem. We tried approximations to solutions of the problem.

The solutions, presented in this work of synthesis, require an understanding of the problem in the context in which it falls, marketing / email marketing, the selection of variables available in the database of the Company, and its analysis.

We applied descriptive techniques, univariate and multivariate inferential techniques.

We used parametric techniques, very few, but we essentially had to use non-parametric techniques. These are presented and explained in this report and, of course, applied to the study.

Keywords: email marketing, chi-square normality, Kruskal-Wallis test, multiple comparisons, correspondence analysis.

Conteúdo

Índice de figuras

Índice de tabelas

Termos mais utilizados em email marketing

A linguagem utilizada na área da informática em conjugação com a área de marketing e, em particular, no email marketing (publicidade enviada através de mensagens eletrónicas) é muito específica. Existem, assim, alguns termos que necessitam ser definidos. [9] [10]

Nesse sentido, e como alguns destes termos são bastante utilizados ao longo deste trabalho, segue-se uma breve lista de conceitos associados ao email marketing.

Os conceitos encontram-se apresentados por ordem alfabética. Por este facto, em algumas definições, são usados termos ainda não definidos.

Ad

Anúncio que aparece numa página da Web projetado de forma a que o utilizador clique sobre mesmo para obter mais informações.

Banner Publicitário (Banner)

Imagem publicitária colocada numa página Web que normalmente funciona como ligação para a página do anunciante.

Blocked

Aviso de bloqueado. Significa que a mensagem não passou, isto é, não foi entregue por ser considerada spam. Isto pode acontecer por estar numa lista proibida ou por ser de um domínio que esteja bloqueado.

Campaign (Campanha)

A campanha é uma mensagem que permite promover algum bem ou serviço. É enviada para um grupo específico de destinatários – indivíduos pertencentes a uma, ou várias listas de subscritores.

Click Through

Corresponde ao número de vezes que se clicou na hiperligação da mensagem (Banner). Muitas vezes é referido como CTR (Click Through Rate). No entanto

este termo indica a razão entre o número de vezes que se clicou sobre o número de visualizações.

Endereço IP

O endereço de IP (Internet Protocol) é um endereço único que permite localizar um computador. Este endereço é uma sequência de números composta por quatro octetos separados por ponto. Um octeto ou byte é uma sequência de oito bites, que é a menor unidade de informação processada por um computador e toma apenas os valores 0 ou 1. A divisão do endereço em octetos facilita a organização da rede. Os dois primeiros octetos são referentes à rede e os dois últimos referentes à identificação do computador.

False Positives

Mensagens rotuladas como spam sem que o sejam. Pode causar prejuízos enormes às empresas. Tem que haver especial cuidado com esta situação.

Marketing

Como conceito é difícil de definir. De facto existem muitas definições técnicas para Marketing.

Mas Marketing é mais que um conceito. É uma ciência ou teoria dotada de ferramentas próprias que são aplicadas em estratégias de negócios, tentando atingir os objetivos das empresas e, em termos muito simplistas, visa o aumento da satisfação dos consumidores e sua consequente fidelização.

Newsletter

Boletim informativo que contém informação sobre a atividade ou serviços de uma organização, empresa ou outra entidade, enviado por correio eletrónico aos seus subscritores.

Open Rate

É a proporção do número total de emails abertos, sobre o total de emails enviados.

Opt-In-list

Uma condição necessária para usar email marketing é ser proprietário de listas de subscritores dos seus serviços. Estas listas contêm informação variada acerca dos indivíduos que as subscreveram assim como os endereços de email.

As listas de subscritores são designadas por listas opt-in, ou subscriber list. De facto nestas listas só constam os utilizadores que decidiram receber emails publicitários.

Spam

Mensagem de email que não é desejada pelo destinatário.

Subscriber

Indivíduo que se inscreve num determinado sítio da Internet para receber mensagens de uma determinada empresa ou entidade.

Web (Worl Wide Web)

É um sistema de servidores de Internet, que suporta documentos formatados, designados por páginas e/ou sites.

Nem todos os servidores de Internet fazem parte da Web. O acesso Web faz-se através de aplicações a que se chamam navegadores (browsers).

Web designer

É o profissional responsável pela elaboração do projeto estético e funcional de uma página Web.

Web marketing

Utilização adequada de recursos, ações e estratégias de marketing via Web.

1. Introdução

Este relatório é relativo ao desenvolvimento de um projeto numa empresa de web marketing sediada no Porto. Este projeto enquadra-se na unidade curricular “Estágio Curricular” do Curso de Mestrado em Estatística de Sistemas, perfil Engenharia e Estatística.

Em todo o trabalho, a empresa de web marketing onde teve lugar o Estágio, decorrido entre 3 de Janeiro de 2011 e 30 de Junho de 2011, é identificada como Empresa.

Este relatório está organizado em seis capítulos como se descreve a seguir. Neste primeiro capítulo apresentamos a Empresa, definimos o problema e fazemos o seu enquadramento tendo em conta o objetivo que se pretende atingir. Falamos ainda, muito superficialmente, da metodologia adotada para a resolução do problema.

No segundo capítulo é feita a contextualização do trabalho, referindo-se brevemente a história da Internet, até ao aparecimento da publicidade por correio eletrónico. Neste tema insere-se uma das principais atividades da Empresa que é novamente descrita, com um pouco mais de detalhe, neste capítulo.

No terceiro capítulo são apresentados alguns conceitos teóricos relativos à metodologia utilizada. Neste capítulo faz-se uma breve explicação da análise de variância não paramétrica, bem como dos testes de independência do qui-quadrado e da análise de correspondências.

No quarto capítulo são apresentados os principais resultados obtidos da análise dos dados em estudo. Os dados foram divididos em dois conjuntos, subscritores com interesse e subscritores sem interesse, conforme estes revelaram ou não interesse por campanhas publicitárias por correio eletrónico, apresentando-se os resultados referentes aos primeiros.

No quinto e sexto capítulo são apresentadas as principais conclusões do estudo bem como sugestões de melhoria futura para a empresa.

1.1 A Empresa e o projeto

A Empresa iniciou a sua atividade no mercado português em 2007 tendo, a partir de 2008, expandido o seu negócio para os mercados espanhol, francês e italiano, brasileiro, polaco e da África do Sul.

Atualmente a Empresa conta com já 30 colaboradores e tem uma sucursal no Brasil. Encontra-se, portanto, em forte expansão.

Dedica-se essencialmente à publicidade online mas não só. Visa gerar negócio para os seus clientes através de soluções de marketing digital orientada aos resultados, conforme se pode ler na sua página publicitária na Web.[16]

Neste trabalho vamos cingir-nos ao tema email marketing, forma de publicidade online, analisando dados fornecidos pela Empresa com vista a identificar os factores facilitadores do consumo de email marketing.

Estes dados, extraídos da informação constante nas listas de subscritores de que a Empresa é detentora, depressa se revelaram escassos (no sentido da informação útil que comportam) e mal organizados.

Quando é enviada uma campanha por email, esta publicita vários produtos. O objetivo é que o recetor do email (subscritor) o abra, clique no banner e compre ou subscreva o produto, ou um dos produtos anunciados.

Este procedimento, muito importante para o nosso trabalho e para a Empresa, designa-se conversão.

É importante ter consciência que a Empresa é remunerada de acordo com o número de conversões por campanha. A Empresa contabiliza o número de conversões que um envio, ou seja, que uma campanha obtém e recebe de acordo com esse valor.

Neste momento não é possível a Empresa identificar que subscritor ou mesmo que lista de subscritores mais converteu. É apenas possível contar os

cliques (ver Click Through) efetuados que são, naturalmente, um indicador para as conversões.

Relativamente às conversões por campanha não nos foi fornecida essa informação. Assim, não temos dados relativos a conversões. Os dados que dispomos indicam-nos, e apenas em alguns casos, se o recetor do email visionou a campanha e clicou no banner.

De qualquer maneira, a Empresa manifestou explicitamente interesse em conhecer melhor o consumidor desta forma de marketing.

Foi proposto que se estudasse, tendo em conta a informação disponível, o que facilitaria o consumo deste tipo de produto

Com este estudo a Empresa pretende saber (com mais rigor) como minimizar o envio de campanhas a subscritores (ou classes de subscritores) que não visionam os emails e, simultaneamente, como maximizar o envio de campanhas àqueles que estão, de facto, potencialmente interessados.

Como referimos, o Estágio curricular decorreu entre 2 de Janeiro e 30 de Junho de 2011. Nesse período tentamos perceber como responder ao desafio lançado com os dados constantes nas enormes, mas desorganizadas bases de dados existentes naquela Empresa.

Este desafio revelou-se cada vez mais ambicioso à medida que o tempo decorria.

1.2 Metodologia

Interessou-nos, em primeiro lugar, perceber como se faz uma campanha publicitária através de correio eletrónico, isto é, como funciona o email marketing, conhecer a “filosofia” de marketing da empresa e adaptarmo-nos à linguagem específica da área e da Empresa.

A Empresa é constituída pelos departamentos: de marketing, informática, e e web design.

Como já foi referido anteriormente, para avaliar se um indivíduo consome ou não publicidade recebida via correio eletrónico precisamos de conhecer o número de mensagens publicitárias que este converte, ou seja o número de conversões.

O número de conversões é, assim, o indicador absoluto do sucesso ou insucesso de uma determinada campanha. No entanto, como também já foi referido, este número não é conhecido.

Usamos, então, um outro indicador, que pode ser útil para obter um valor próximo do número de conversões, o número de cliques. Como já dissemos, mas recordamos agora devido à importância do termo no nosso trabalho, clicar é o ato de carregar no banner ou hiperligação que abre a página do produto ou serviço que está a ser publicitado.

Recorrendo a técnicas da Estatística o objetivo é isolar os factores que influenciam o aumento ou diminuição do número de cliques.

- **Variáveis disponíveis para o estudo**

As variáveis a que temos acesso, para cada subscritor, são o género, a idade, a localização geográfica, o domínio da caixa de correio eletrónica e a lista de subscrição dos indivíduos.

Importa referir que existem listas de subscritores em que não temos toda esta informação.

Convém também salientar que a localização geográfica foi obtida através de um algoritmo que cruza a informação fornecida pelo utilizador com a informação relativa ao endereço de IP. O endereço de IP está, relacionado com o local onde um utilizador acede à Internet que pode variar ao longo do tempo. Este facto torna esta variável pouco fidedigna, pelo que optamos por ignorá-la neste projeto. No entanto pensamos que a localização geográfica do subscritor é de grande importância para este tipo de estudo.

Das técnicas estatísticas usadas, além da Análise Descritiva e Exploratória de Dados, destacamos a Análise de Variância com comparações múltiplas, Testes de Independência do Qui- quadrado e Análise de Correspondências.

A utilização destas técnicas foram apoiadas computacionalmente recorrendo aos softwares R 2.13.3 e Excel 2007.

2. Contextualização

Para melhor percebermos este tipo de empresas e área de negócio onde estão incluídas, vamos explorar ainda:

1) Conceitos associados ao email marketing uma vez que é de grande importância familiarizarmo-nos com a linguagem associada à Internet e ao marketing.
2) O surgimento e as condições que permitiram o aparecimento do email marketing.
3) O enquadramento da Empresa na área de negócio, tentando compreender, ainda que sem grandes ambições, a filosofia que lhe está subjacente assim como a sua metodologia de trabalho.

2.1. Um pouco de História: a Internet

Durante Após a segunda guerra mundial, em 1945, os governantes dos países aliados reuniram-se para estabelecer regras de divisão do território alemão bem como de divisão do território da Europa de Leste. Esta reunião conduziu à Guerra Fria, já que a Rússia passou de aliado a inimigo. Esta guerra era basicamente constituída por ações de espionagem de ambas as partes, aliados e russos, com o objetivo de impedir o ataque do adversário.

Em contrapartida a Guerra Fria proporcionou grande evolução e desenvolvimento, quer no que diz respeito à tecnologia, quer no que diz respeito à ciência.

Em Outubro de 1957 a Rússia lançou para o espaço o primeiro satélite artificial, Sputnik. Em resposta a este avanço, os americanos criaram a Advanced Research Project Agency (ARPA), cujo principal objetivo era o desenvolvimento de programas espaciais.

Em 1961 a Universidade da Califórnia herdou um computador e a investigação da ARPA foi conduzida para a área da informática, iniciando a sua investigação no desenvolvimento de redes de comunicação de dados. Um dos

seus investigadores, Licklider criou um sistema de comunicação interativa de transmissão de dados. No entanto sentiu-se a necessidade de existência de uma rede-NET que possibilitasse comunicações rápidas. Este desafio que surgiu não era de fácil resolução pois existiam diversas redes, cada uma com as suas regras e linguagem incompatíveis com as outras.

A fiabilidade da rede e a confiança do utilizador eram outro problema que se colocava, mas Robert Taylor, sucessor de Licklider nos estudos da ARPA, propôs uma solução para esta questão. Esta solução contemplava o uso de redes do tipo distribuído nas quais era possível a conexão de emissor e recetor utilizando diferentes caminhos da rede. Caso alguma ligação ou nó da rede avariasse, a mensagem correria naturalmente por um caminho alternativo.

Paul Baran e Donald Davies, investigadores da área, asseguraram a questão da segurança das mensagens com um sistema em que as mensagens nunca circulavam completas, ou seja as mensagens seguiam em fragmentos até ao computador recetor que reconstruiria a mensagem original. Contudo esta ideia aumentava o trabalho dos computadores emissor e recetor, pelo que foram construídos computadores intermediários que processavam o trabalho.

Com estes avanços, no fim da década de 1960, nasceu a primeira rede de computadores, ARPANET. Esta rede foi criada entre a Universidade da Califórnia, Universidade de Utah e o Standford Research Institute, ficando para a história o dia 1 de Dezembro de 1969 como o dia do seu surgimento. Os nós de rede iniciais rapidamente foram alargados e, em simultâneo, eram construídas outras redes nos Estados Unidos, bem como em alguns países da Europa entre estes Portugal.

Na década de 1970 uma equipa de investigadores liderada por Vinton Cerf e Robert Kahn desenvolveu um protocolo, denominado TCP/IP (Transmission Control Protocol and Internet Protocol) que garantia a conexão entre todas as redes internacionais, regionais e nacionais, que até então não comunicavam entre si. Em 1977, os dois investigadores realizaram uma demonstração do TCP/IP, considerando-se que nesse momento nasceu a Internet.

Assim, a Internet é um conjunto de redes interligadas que estão regidas pelo mesmo protocolo.

Em meados dos anos 80 a Internet começou a difundir-se em Portugal sendo utilizada por algumas universidades e empresas, chegando a todas as universidades no início dos anos 90. Em 1995, os meios de comunicação social divulgaram a existência e utilidade da Internet, dando-se uma explosão na utilização deste novo meio de comunicação [12].

Hoje em dia, a Internet apresenta-se como o meio de comunicação por excelência, contribuindo para uma nova maneira de ver e estar no mundo. Podemos falar no mundo antes do surgimento da Internet e no mundo depois do surgimento da Internet (aldeia global).

Parece por demais evidente as empresas usarem e explorarem, em crescendo, este meio. Seja para atingirem novos públicos, ou para prestarem apoio aos seus clientes.

Nos países desenvolvidos a Internet já substitui muitos serviços, públicos ou privados, evitando idas a bancos, ao supermercado, compra de livros, de roupa, etc. É um sem fim de mordomias que a Internet nos proporciona. Resta ponderar o risco associado a cada um destes serviços.

A criptografia está em franco desenvolvimento com o objetivo de tornar a Internet mais segura. No entanto, todos sabemos que havendo um código há sempre uma chave que o decifra. Este problema é naturalmente de outro âmbito mas não nos pode ser indiferente quando falamos de Internet e dos seus atributos.

2.2. Web marketing e o Email marketing

O Web marketing é uma forma de publicidade através da Internet. Passou a ser um meio privilegiado para a divulgação/promoção de produtos e serviços. O objetivo da publicidade é divulgar informação útil (tendo em vista o propósito que pretende atingir) sobre um produto ou serviço, influenciando, ou não, o seu

consumo. É preciso não esquecer que existe publicidade que tem fins “não comerciais”. Neste caso pretende-se informar e divulgar, tendo em vista o bem estar do cidadão.

Despertar no consumidor a atenção, o interesse, o desejo, a memorização e a ação (AIDMA) são as etapas da publicidade. [8]

Neste trabalho estamos a falar de publicidade comercial cujo objetivo último é levar o cidadão a consumir. A publicidade online, de que falamos neste trabalho, tem um objetivo comercial.

A enorme vantagem relativamente a outros tipos de publicidade (seja através dos media, cartazes de divulgação, marcas no vestuário que usamos, nos sacos de supermercado, etc.) é de estar a um clique de distância do consumidor. Basta ao consumidor clicar na peça para obter informação detalhada sobre o artigo ou serviço publicitado, comparar preços e até efetuar a sua subscrição no site ou mesmo efetuar uma compra. [11]

Referimos já por diversas vezes que a forma de publicidade na Internet que nos interessa é o email marketing. É extremamente difícil precisar o momento em que se começou a utilizar o correio eletrónico como forma de fazer chegar a publicidade aos consumidores.

De acordo com Fernando Zamith, da agência Lusa, o primeiro email foi escrito há mais de quarenta anos e foi enviado dois meses após o primeiro nó de Internet e contendo apenas “LO.”. O emissor desta mensagem, Leonard Kleinrock da Universidade da Califórnia em Los Angels, pretendia escrever “LOGIN”, mas o sistema foi abaixo, pelo que o recetor, Douglas Engelbart do Stanford Research Institute, não a recebeu completa.

Entretanto foram surgindo novos sistemas de troca de mensagens e no início da década de 1970, Ray Tomlinson inventou os primeiros programas de envio de mensagens de correio eletrónico. Rapidamente este se tornou uma ferramenta de trabalho, um meio de comunicação e portanto uma ferramenta de publicidade. [3]

Com a expansão da Internet e o uso do correio eletrónico, muito rapidamente as empresas perceberam que este seria um novo meio de publicitar e vender os seus produtos, mais rápido e, provavelmente, bastante mais económico que o sistema que usavam até então. [4]

Claro que existem regras para fazer publicidade via Internet. O email marketing assenta na permissão do utilizador, uma vez que este, quando recebe uma mensagem publicitária, tem que ter permitido o seu envio. A permissão de envio destas mensagens é feita através da opção de as receber num formulário de um website ou da escolha da receção de uma newsletter. Qualquer mensagem publicitária enviada sem permissão é considerada spam. [11]

O envio de spam foi um dos entraves à difusão do email marketing, pois ambos eram confundidos, inclusive pelos filtros de spam que bloqueavam tudo que se assemelhasse com este. Mas cedo os profissionais desta área se adaptaram às adversidades e descobriram técnicas que facilitassem a entrada dos seus emails nas caixas dos seus clientes. No presente, o email marketing é essencial na forma de comunicação e contacto entre empresas e clientes. [4]

Regido pelo principio ético de só enviar emails para quem os solicita, o email marketing demarca-se do spam e constitui uma mais valia para a oferta ou promoção de artigos ou serviços, bem como de campanhas de fidelização entre outros. Este tipo de publicidade também é usado para envio de informações periódicas, noticias, artigos, comunicados internos e até envio de convites para eventos. Assim o email marketing constitui uma forma simples de contactar e ser lembrado, pois permite o envio regular de mensagens direcionadas e obter rapidamente a respetiva resposta. Esta é a principal vantagem do email marketing, além da interatividade que permite com o cliente, a segmentação e personalização das campanhas e a mensuração imediata dos resultados por parte da empresa publicitária. [18]

2.3. A Empresa

A Empresa onde se realizou o estágio dedica-se a vários serviços online sendo a publicidade um dos seus pontos fortes. Esta empresa apresenta estratégias e soluções de acordo com o produto que se pretende vender, bem como de acordo com as expetativas e objetivos do cliente que lhe solicita o serviço.

Como já foi referido, a Empresa deu os primeiros passos trabalhando no mercado nacional e contando com três colaboradores. Em 2008 atravessou a fronteira e com cinco colaboradores e entrou no mercado espanhol. No ano seguinte, já com mais do dobro dos colaboradores penetrou no mercado francês e italiano. No início da segunda década do novo milénio, com duas dezenas de colaboradores, chegou aos mercados brasileiro, polaco e da África do Sul. Atualmente mais de trinta colaboradores possibilitam a sua forte presença no mundo e a franca expansão e crescimento.

Do que senti e pude constatar enquanto estagiária desta empresa de marketing digital foi que esta se rege por princípios de criatividade, sustentabilidade, paixão e, como não podia deixar de ser, faz sobressair a relevância do produto de cada um dos seus clientes - que são tratados como únicos.

O que é um facto é que a Empresa revela elevada performance no seu meio pois parece conseguir estar sempre um passo à frente das suas empresas concorrentes.

Enquanto outras empresas se limitam a publicitar produtos, esta pretende interagir com o consumidor e, sobretudo, vender o produto que anuncia assumindo as despesas da geração das vendas, pagando ao cliente apenas os resultados que recebe. Evidentemente que a Empresa pretende não só aumentar o número de vendas como também ser responsável pela visibilidade e promoção de uma marca neste mundo global que é a Internet.

A Empresa pretende mostrar a capacidade de gerar negócio e construir um futuro de sucesso com os seus parceiros. [16]

Figura 1: Países de actuação da empresa

3. Técnicas de suporte ao desenvolvimento do trabalho: Estatística

Depois do conhecimento da Empresa, e do meio em que se insere, é importante o conhecimento das técnicas estatísticas que utilizamos para tentar responder às questões colocadas de forma a ajudar a melhorar os seus resultados.

Usamos essencialmente técnicas descritivas com o objetivo de identificar e/ou traçar perfis dos consumidores de publicidade por correio eletrónico.

Além da análise descritiva univariada dos dados, realizamos análises de variância para identificar diferenças entre grupos de indivíduos (inferindo para os valores centrais das populações) e, quando estas existem, realizam-se comparações múltiplas com o objetivo de identificar os responsáveis pelas diferenças.

A análise de variância utilizada neste estudo é a não clássica devido à ausência de normalidade dos dados.

Efetuamos testes de independência do Qui-quadrado com o objetivo de conhecer a existência de associação (ou não) entre atributos na mesma população.

Com o intuito de traçar perfis do utilizador de email marketing aplicamos técnicas de análise de correspondências.

3.1. Análise da Variância

A análise da variância clássica, mais conhecida por ANOVA (ANalysis Of VAriance), tem como base pressupostos de normalidade e homocedasticidade dos dados, além da independência das populações em estudo.

O objetivo desta análise é avaliar a igualdade dos valores médios das diferentes populações, quando estas apenas podem diferir no seu valor médio. As populações são supostamente normais e todas devem ter a mesma variância.

O estudo da normalidade dos dados pode ser feito usando diferentes metodologias. Algumas gráficas tais como construção de histogramas (função densidade de probabilidade empírica) ou gráficos de quantis (que compara os quantis teóricos com os quantis empíricos), ou através do cálculo de medidas de achatamento e curtose e/ou através de testes de hipóteses. Existe um elevado número de testes de normalidade, muitos deles implementados no sistema computacional R.

Talvez o teste mais usado na literatura seja o de Kolmogorov-Smirnov pois é dos mais conhecidos, embora não seja especialmente pensado para testar normalidade. De qualquer maneira alguns dos testes de normalidade derivam deste. Para a normalidade, alguns testam somente a simetria e a curtose (um de cada vez ou ambos simultaneamente) tendo como base as bem conhecidas distribuições das estatísticas usualmente designadas por b_3 e b_4 (assimetria e

achatamento). Estas estatísticas são definidas à custa dos momentos centrais de ordem três e quatro respetivamente. O momento de ordem r, para uma amostra de dimensão n, é dado por

$$m_r = \frac{1}{n}\sum (X_i - \bar{X})^r$$

Desta forma, os coeficientes de assimetria e achatamentos são expressos pelas seguintes fórmulas

$$b_3 = \frac{m_3}{(\sqrt{m_2})^3} \qquad \text{e} \quad b_4 = \frac{m_4}{(\sqrt{m_2})^4}$$

O teste de Shapiro-Wilk é um dos testes de normalidade mais poderosos, especialmente para amostras pequenas. A normalidade é testada por comparação de duas alternativas para a estimação da variância, sendo uma delas não paramétrica. [13]

O teste de Jarque-Bera é muito usado em Econometria. Baseia-se nas medidas de simetria e curtose e na distribuição assintótica de b_3 e b_4 que, supondo a hipótese nula verdadeira (ou seja a normalidade dos dados), tem distribuição Qui-quadrado com 2 graus de liberdade. [2]

A package nortest do sistema computacional R contém 5 testes de normalidade: Shapiro-Francia, Anderson-Darling, Cramer-Von Mises, Qui-quadrado de Pearson's e Lilliefors.

O estudo da homocedasticidade ou homogeneidade de variâncias é classicamente efetuado com o teste de Bartlett [17]. No entanto este teste é sensível à não normalidade das populações. O teste de Levene é uma alternativa a este pois não é tão sensível à ausência de normalidade.

Assim, quando os pressupostos da ANOVA (clássica) não se verificam devemos usar uma alternativa não paramétrica como, por exemplo, o teste de Kruskal-Wallis.

Tal como na ANOVA, com este teste pretendemos decidir se k populações independentes são idênticas a menos da sua localização central. A única

suposição exigida por esta prova é a de que a variável em estudo seja contínua e tenha mensuração no mínimo ao nível ordinal. Como usamos este teste no presente trabalho, vamos tecer mais considerações acerca do mesmo.

3.2. Teste de Kruskal-Wallis

O teste de Kruskal-Wallis é o análogo ao teste F utilizado na ANOVA a 1 factor.

Figura 2: William Henry Kruskal (1919 - 2005)

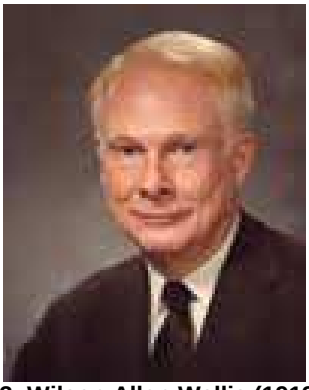


Figura 3: Wilson Allen Wallis (1912-1998)

A construção do teste passa, em primeiro lugar, por atribuir uma ordem, ou rank (graduação) a cada uma das observações.

A ordem 1 é atribuída à menor das observações, a ordem 2 à segunda menor e assim sucessivamente. Caso existam repetições, atribuímos a média das ordens que atribuiríamos caso não existissem. Claro que se temos n observações, a ordem n é atribuída à maior das observações.

Em seguida somam-se as ordens para cada amostra e calcula-se o total destas somas parciais.

Através da prova de Kruskal-Wallis analisa-se se esses totais são tão distintos que não possam ser considerados da mesma população.

Para analisar a existência da disparidade atrás referida, caso as amostras não sejam demasiado pequenas, usa-se a estatística H que tem distribuição Qui-quadrado com $k - 1$ graus de liberdade.

$$H = \frac{12}{N\,(\,N+1\,)}\sum_{j=1}^{k}\frac{R_j^2}{n_j} - \,3\,(N+1)$$

onde k representa o número de amostras, n_j o número de casos na amostra j, $N = \sum n_j$ é o número de casos em todas as amostras combinadas, R_j é a soma das ordens na amostra j e $\sum_{j=1}^{k}$ indica o somatório sobre todas as k amostras.

Por último verifica-se se o valor da probabilidade associado a H não excede o nível de significância estipulado. Neste caso rejeita-se a hipótese das amostras serem provenientes da mesma população, no sentido em que diferem na localização central.

Segundo Andrews, em condições de aplicabilidade do teste paramétrico mais poderoso, o teste de análise de variância através das ordens, como também é conhecido este teste, revela uma eficiência assintótica de 95,5%. [15] [19]

No caso da rejeição da hipótese de igualdade é necessário identificar os responsáveis por tal rejeição realizando comparações múltiplas.

Nesta situação, as comparações múltiplas são aplicadas à soma das ordens, avaliando as suas diferenças duas a duas. As somas das ordens são, assim, comparadas aos pares e verifica-se se a sua diferença é ou não significativa. Caso esta diferença seja significativa, então o par em causa é responsável pela rejeição da hipótese nula, caso contrário não o é.

3.3. Teste de independência do Qui-quadrado

O teste do Qui-quadrado surge em vários contextos devido às suas diversas aplicações.

Vulgarmente este é usado como teste de ajustamento, teste de homogeneidade e teste de independência. É efetuado sobre tabelas de contingência.

No nosso trabalho interessa-nos perceber se existe ou não associação entre pares de algumas das variáveis estudadas.

O teste do Qui-quadrado é aplicado a duas amostras classificadas em I e J categorias, respetivamente.

Consideramos uma amostra classificada segundo dois critérios ou atributos categóricos A e B. Pretendemos concluir se estes dois atributos duma mesma população são independentes entre si, ou se há alguma associação entre eles.

Suponhamos que existem n observações independentes em n indivíduos que estão agrupados em $I \times J$ grupos. Estes grupos são formados de acordo com os critérios A e B, que possuem respetivamente I e J categorias.

Na tabela 1 estão representadas as frequências observadas em cada grupo, v_{ij}, onde i a i-ésima categoria do atributo A e j a j-ésima categoria do atributo B, com $i = 1, \dots, I$ e $j = 1, \dots, J$.

	B				
A	1	2	...	J	**Total**
1	v_{11}	v_{12}	...	v_{1J}	$\boldsymbol{v_{1\cdot}}$
2	v_{21}	v_{22}	...	v_{2J}	$\boldsymbol{v_{2\cdot}}$
⋮	⋮	⋮		⋮	⋮
I	v_{I1}	v_{I2}	...	v_{IJ}	$\boldsymbol{v_{I\cdot}}$
Total	$\boldsymbol{v_{\cdot 1}}$	$\boldsymbol{v_{\cdot 2}}$	...	$\boldsymbol{v_{\cdot J}}$	$\boldsymbol{n}$

Tabela 1:Tabela de contingência

Para avaliar a independência entre os dois critérios ou atributos recorremos à definição de acontecimentos independentes.

Sabemos que dois acontecimentos X e Z são independentes se a probabilidade da interseção dos dois for igual ao produto das probabilidades de cada um. Ou seja, X e Z dizem-se independentes se verificarem a igualdade

$$P(X \cap Z) = P(X) \cdot P(Z).$$

Assim, a hipótese nula H_0 que se pretende testar é equivalente à igualdade

$$p_{ij} = p_{i\cdot}p_{\cdot j}$$

com $i = 1, \dots, I$ e $j = 1, \dots, J$. Como habitualmente as probabilidades marginais são dadas por:

$$p_{i\cdot} = \sum_{j=1}^{J} p_{ij} \quad \text{e} \quad p_{\cdot j} = \sum_{i=1}^{I} p_{ij}$$

as probabilidades $\mathrm{p_{i\cdot}}$ e $\mathrm{p_{\cdot j}}$ são, respetivamente, a probabilidade da categoria i e da categoria j.

Assumindo que todas as probabilidades são desconhecidas, temos que estimar $\mathrm{I} + \mathrm{J} - 2$ parâmetros, ou seja

$$p_{i\cdot} \text{ para } i = 1, \dots, I-1 \text{ e } p_{\cdot j} \text{ para } j = 1, \dots, J-1$$

uma vez que,

$$p_{I\cdot} = 1 - \sum_{i=1}^{I-1} p_{i\cdot} \quad \text{e} \quad p_{\cdot J} = 1 - \sum_{j=1}^{J-1} p_{\cdot j}$$

Para estimar tais parâmetros recorre-se à função de verosimilhança $\ell(\boldsymbol{p})$ para o vetor das frequências $\boldsymbol{v} = [v_{11} \ \dots \ v_{IJ}]^T$

$$\ell(\boldsymbol{p}) = \prod_{i,j} {p_{ij}}^{v_{ij}} \qquad \text{com} \qquad \boldsymbol{p} = [p_{11} \dots p_{IJ}]^T$$

Sob a hipótese nula tem-se,

$$\ell(\boldsymbol{p}) = \prod_{i,j} {p_{i\cdot}}^{v_{ij}} \, {p_{\cdot j}}^{v_{ij}}$$

Deste modo, aplicando logaritmos, como usualmente, vem que

$$\ln \ell(\boldsymbol{p}) = \ln\left(\prod_{i,j} {p_{i\cdot}}^{v_{ij}} {p_{\cdot j}}^{v_{ij}}\right)$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} v_{ij}\left(\ln p_{i\cdot} + \ln p_{\cdot j}\right)$$

$$= \sum_{i=1}^{I} v_{i\cdot} \ln p_{i\cdot} + \sum_{j=1}^{J} v_{\cdot j} \ln p_{\cdot j}$$

onde

$$v_{i\cdot} = \sum_{j=1}^{J} v_{ij} \quad \text{e} \quad v_{\cdot j} = \sum_{i=1}^{I} v_{ij}.$$

Usando a equação acima facilmente se calculam os estimadores de máxima verosimilhança (através da resolução do sistema associado às derivadas parciais de $\ln \ell(\boldsymbol{p})$), cujas soluções $\widehat{p_{i\cdot}}$, $\widehat{p_{\cdot j}}$ de $p_{i\cdot}$, $p_{\cdot j}$ são dadas por

$$\widehat{p_{i\cdot}} = \frac{v_{i\cdot}}{n}, \qquad \widehat{p_{\cdot j}} = \frac{v_{\cdot j}}{n}, \qquad i = 1, \dots, I; \quad j = 1, \dots, J.$$

Daqui obtemos $\widehat{p_{ij}}$, estimador de máxima verosimilhança de p_{ij},

$$\widehat{p_{ij}} = \widehat{p_{i\cdot}}\,\widehat{p_{\cdot j}} = \frac{v_{i\cdot}}{n}\frac{v_{\cdot j}}{n}$$

Desta forma temos

$$\chi_n^2(\boldsymbol{p}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(v_{ij} - np_{ij})^2}{np_{ij}},$$

Assim como,

$$\chi_n^2(\widehat{\boldsymbol{p}}) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(v_{ij} - n\widehat{p_{ij}})^2}{n\widehat{p_{ij}}}$$

A regra de decisão para este teste diz que H_0 é rejeitada se $\chi_n^2(\widehat{\boldsymbol{p}}) > c_\alpha$, onde c_α representa o quantil α da distribuição χ^2 com $(I-1)(J-1)$ graus de liberdade.

No caso da tabela 1 se resumir a uma tabela com duas categorias em cada critério, estamos perante o caso particular de $I = J = 2$. Neste caso a estatística de teste é menos complicada de calcular

$$\chi_n^2(\hat{\boldsymbol{p}}) = n\frac{(v_{11}v_{22} - v_{12}v_{21})^2}{v_{1\cdot}v_{2\cdot}v_{\cdot 1}v_{\cdot 2}}$$

e sob H_0 quando a dimensão da amostra é muito grande esta estatística aproxima-se de χ_1^2. [6]

Limitações do teste: O teste de independência do Qui-quadrado é muito útil, mas nem sempre pode ser utilizado.

Este teste só deve ser utilizado quando menos de um quinto das células têm frequência esperada inferior a cinco e quando todas as células têm a mesma não inferior a um, pois caso contrário existiriam frequências nulas (ou próximas de zero) que obrigavam à divisão por zero, o que não é possível.

Se estas condições não forem satisfeitas, as categorias devem ser agrupadas de forma a tornar os valores esperados de frequência em valores passíveis de aplicação deste teste.

Por vezes as tabelas de contingência têm categorias cujo valor observado é nulo, ou seja não verificam as condições de aplicabilidade do teste de independência do qui-quadrado. Estas tabelas são consideradas tabelas de contingência especiais.

As entradas nulas destas tabelas são designadas por zeros estruturais e estas tabelas denominam-se incompletas. A análise de tais tabelas constitui um caso particular das tabelas de contingência e foram muitos os autores que se dedicaram a este problema, nomeadamente Goodman (1968), Fienberg e Bishop (1969) e Mantel (1970).

Uma forma de realizar o estudo das tabelas de contingência incompletas consiste em eliminar as células nulas substituindo-as por uma constante maior que zero, por exemplo 0.5 e, posteriormente, executar o estudo pretendido.

Fienberg apresentou um método mais formal para determinar a constante que substituiria os zeros da tabela.

Obviamente que este método deixa de ser prático e exequível no caso de a tabela ter muitos zeros estruturais. [5]

3.4. Análise de correspondências

A análise de tabelas de contingência é um método bastante importante na análise multivariada de dados.

O termo contingency parece ter sido usado pela primeira vez por Karl Pearson (1904) para descrever o desvio relativamente à independência entre linhas e colunas de uma estrutura de dados disposta numa tabela.

Mais recentemente o termo refere-se a contagens (frequências). Consequentemente uma tabela de contingência contém informação de natureza discreta ou categórica.

O desenvolvimento de técnicas adaptadas a problemas que envolvem tabelas de contingência são devidas fundamentalmente a Karl Pearson, G. Undy Yule e R. A. Fisher. [7]

Uma das mais importantes medidas de associação entre duas variáveis categóricas é a estatística de Qui-quadrado de Pearson. De facto, Pearson desenvolveu um trabalho fundamental na estatística do Qui-quadrado usando-a para comparar as frequências observadas com as que se esperavam obter sob hipótese de independência entre duas populações. [1]

A metodologia da análise de correspondências (AC) insere-se nas técnicas de redução de dimensionalidade pois o seu principal objetivo é encontrar o espaço de menor dimensão, onde a amostra possa ser bem representada e identificar a existência de algum padrão.

De facto, e em termos simples, o objetivo do método é descrever as linhas e colunas de uma tabela de contingência, ou seja, estudar a dependência entre os indivíduos e as categorias dos atributos em estudo.

Baseia-se na decomposição do Qui-quadrado e o estudo da dependência tem como base representações gráficas. Dois indivíduos ou duas categorias estão tão mais relacionados quanto mais próximos estiverem um do outro. Importa saber qual a definição de proximidade a utilizar.

O ponto de partida é uma tabela de contingência nas quais se tem n indivíduos ou objetos classificados de acordo com m atributos. A informação pode ser agrupada numa matriz A de dimensão $n \times m$.

As entradas da matriz A são números inteiros não negativos. Como habitualmente, mas sem perda de generalidade (em Análise de Correspondências existe simetria entre o estatuto das linhas e colunas da matriz), nas colunas temos as variáveis, nas linhas os indivíduos.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix}$$

Para comparar as diversas linhas da matriz com base nas colunas, ou vice-versa, é necessário transformar os valores em frequências relativas dividindo a matriz A pelo número total de observações, a.

Deste modo obtém-se a matriz de correspondência, P, constituída pelas frequências relativas das observações.

$$\boldsymbol{P} = \frac{\boldsymbol{A}}{a} = \begin{bmatrix} \frac{a_{11}}{a} & \frac{a_{12}}{a} & \dots & \frac{a_{1m}}{a} \\ \frac{a_{21}}{a} & \frac{a_{22}}{a} & \dots & \frac{a_{2m}}{a} \\ \vdots & \vdots & & \vdots \\ \frac{a_{n1}}{a} & \frac{a_{n2}}{a} & \dots & \frac{a_{nm}}{a} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1m} \\ p_{21} & p_{22} & \dots & p_{2m} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \dots & p_{nm} \end{bmatrix}$$

Uma vez que a técnica AC é invariante em relação a a, usamos a matriz P em vez da matriz A.

- **Perfis linha e perfis coluna**

A ideia é agora transformar a tabela de frequências de forma a eliminar a influência das marginais na comparação das linhas (ou colunas) da matriz.

Os perfis linha são obtidos dividindo cada linha pelo respetivo total de linha, ou seja o perfil linha é dado pelo produto $\mathrm{D_r^{-1}P}$, sendo $\mathrm{D_r}$ a matriz diagonal dos pesos das linhas.

Os pesos das linhas são dados pelo quociente entre o total de cada linha e o total das observações, ou seja o peso da linha i é

$$r_i = \frac{a_{i\cdot}}{a}$$

sendo r o vetor dos pesos das linhas.

De forma análoga definimos o perfil coluna. Este é dado por $\mathrm{PD_c^{-1}}$, sendo $\mathrm{D_c}$ a matriz diagonal dos pesos das colunas, os quais são dados por

$$c_j = \frac{a_{\cdot j}}{a}$$

sendo c o vetor dos pesos das colunas.

A informação precedente serve para calcular o **traço de perfis**.

Cada linha e cada coluna das matrizes anteriores têm soma unitária definindo, respetivamente, um perfil de linha ou um perfil de coluna.

- **Nuvens**

Estes perfis definem duas nuvens de pontos, uma para linhas e outra para colunas.

As nuvens são conjuntos de pontos cujas coordenadas são dadas pelos perfis linha ou coluna (conforme o caso). Estas nuvens são o que usualmente se denomina em análise multivariada como centro de gravidade ou centróide. É conhecido da física que o centro de gravidade de um corpo é o ponto onde a resultante das forças é nula. Neste caso o centróide é determinado através da

generalização do conceito de média (ponderada). Desta forma o centróide é o ponto cujas coordenadas correspondem ao perfil médio.

Os perfis serão tanto mais semelhantes quanto menor for a distância que os separa.

Se um perfil difere muito do centróide, o seu ponto é representado longe da origem. Os perfis que se aproximam dos centróides são representados por pontos próximos do centro de gravidade. Deste modo, se todas as categorias possuírem perfis iguais, todos os seus pontos coincidem com o centróide.

- **Distância, proximidade**

Como sabemos, em Matemática, existem várias distâncias. A distância mais usual é a euclidiana, mas não é apropriada a este caso.

De facto a distância euclidiana depende unicamente do quadrado da diferença entre os perfis de cada individuo. Ora, quando esta diferença é reduzida, o quadrado da diferença entre os perfis de cada individuo acentua a sua fraca contribuição para o cálculo da distância global apresentando um efeito contrário quando a diferença é elevada.

Assim, para servir os propósitos da AC, o conceito de proximidade é medido pela distância ponderada pelo inverso da massa (ou peso), denominada por distância do qui-quadrado.

Esta ponderação é essencial para estabilizar estas diferenças, dando maior peso às proporções mais significativas.

Com a utilização desta distância podemos substituir dois indivíduos (ou categorias) semelhantes que o resultado não se altera. Este aspeto revela-se vantajoso, pois indivíduos idênticos não trazem informação nova, o que garante que os resultados se mantenham invariantes.

- **Inércia**

Um conceito importante em AC, análise de correspondências, e que está ligado ao conceito de distância é o conceito de inércia.

O “momento de inércia” de um objeto, conceito fundamental em Mecânica, é a soma do produto da massa pelo quadrado da distância de todas as partículas que constituem o objeto.

No caso da AC, o conceito de inércia está associado à existência de uma nuvem de pontos, cujas coordenadas são determinadas pelos perfis, que totaliza massa um. Esta nuvem tem um centróide (perfil médio), onde se pode definir uma distância entre pontos: a distância do qui-quadrado.

A inércia das linhas é definida como a média ponderada das distâncias dos perfis de linha ao centro de gravidade com ponderações dadas pelos pesos. A inércia das colunas é análoga à das linhas, sendo ambas iguais à inércia total que é o quadrado do coeficiente de contingência de Pearson. [12]

4. Aplicação

4.1. A base de dados

A base de dados da Empresa tem vindo a ser construída com base nas listas de subscritores que a Empresa vai conseguindo ao longo do tempo.

As listas de subscritores são listas que acumulam informação sobre um indivíduo que subscreve um serviço. Estas listas distinguem-se entre si pelo tema e/ou pela aplicação que as originou.

- **Como se criam as listas e que informação contêm**

Por exemplo a lista “apps.facebook.com/polvo-paul”, que já não existe, foi criada através de uma aplicação do facebook na qual as pessoas tinham de deixar pouca informação pessoal para além do seu endereço de correio eletrónico.

A lista “Ncursos.net” foi criada com os dados dos indivíduos que subscreveram um serviço de informação sobre cursos.

Nesta subscrição os indivíduos facultaram mais informação além do email.

No anexo A1 encontra-se uma tabela com as listas envolvidas neste estudo.

Em geral, além do endereço de correio eletrónico, estas listas têm informação sobre o género, idade, código postal e/ou cidade, endereço de IP, domínio de email, browser utilizado, sistema operativo, entre outras.

Ocorre que alguma desta informação não é fiável. Por exemplo, por vezes a cidade e o código postal, quando comparados com o endereço de IP dão informação contraditória. Outras vezes o utilizador insere sequências de letras ou números que não fazem sentido para a informação solicitada, por exemplo insere “2010” na idade.

Assim, não nos foi possível usar todos os dados constantes na base de dados da Empresa. Tivemos que os retirar logo à partida.

Ora estas listas de subscritores constituem os alicerces do email marketing. De facto, é através delas que a Empresa faz campanhas publicitárias.

A Empresa escolhe uma ou várias listas para a qual envia uma mensagem que publicite um produto ou serviço.

A escolha das listas é efetuada tendo em conta informação anterior sobre o sucesso ou insucesso de determinada campanha (de determinado tipo) e socorre-se muito do bom senso de quem as envia. A ideia é tornar esta escolha mais científica, tendo como base um estudo estatístico.

- **Campanhas**

Cada mensagem de correio eletrónico que promove um produto ou serviço é designada por campanha.

As campanhas foram agrupadas (pelo departamento de marketing da Empresa e com a nossa ajuda) de acordo com a temática envolvida.

Criaram-se categorias de campanhas. Por exemplo as campanhas “PT. MasterChef - NA - Descontos@F” e “PT. MasterChef - NA - Farmville@F” inserem-se na categoria de cozinha. As campanhas “PT. Livra - NA – Filmes”, “PT. Livra - NA - Meteorologia” e “PT. Livra - NA – MSN” inserem-se na categoria dos sorteios. No anexo A2.1 encontra-se uma lista com algumas campanhas e a respetiva categorização.

Sempre que neste relatório se fizer referência a “campanha”, queremos referir-nos a um conjunto de campanhas da mesma categoria.

As categorias estão codificadas por números de um a quarenta e um. Esta codificação encontra-se no anexo A2.2.

A base de dados da empresa contém ainda informação sobre as campanhas enviadas para cada subscritor, bem como sobre o número de aberturas e cliques desse subscritor na campanha.

Importa referir que o contato com a base de dados real da empresa nunca foi possível devido ao desconhecimento do formato em que esta se encontra armazenada.

A Empresa forneceu-nos os dados que considerou serem os melhores. Tratavam-se dos dados que à data eram os mais recentes e tinham mais informação útil sobre os subscritores.

Os dados em estudo são referentes às campanhas enviadas para as listas de Portugal -via mailer Oempro, desde o dia 22 de Outubro de 2010 até ao dia 15 de Janeiro de 2011.

- **Variáveis disponíveis**

Estes dados foram extraídos para uma base de dados em Excel onde constam as variáveis: endereço de correio eletrónico, idade, género, lista ou listas de subscrição, número de emails enviados, número de emails abertos e número de emails clicados de determinado tipo de campanha, domínio do endereço de correio eletrónico e localização geográfica. Na Figura seguinte apresenta-se um excerto da base de dados.

	A	B	C	D	E
1	*email*	*list*	*gender*	*age*	*postalcode*
43		92/	M	24	NULL
44		111/	M	24	8000-354
45		124/	F	40	NULL
46		92/124/	M	18	NULL
47		47/	NULL	NULL	NULL
48		92/	NULL	NULL	NULL
49		124/	M	20	NULL
60		56/	NULL	NULL	NULL

	F	G	H	I
1	*city*	*geo_user*	*geo_ip*	*domain*
43	NULL	pt	PT	letras.up.pt
44	faro	pt	pt	sapo.pt
45	NULL	pt	pt	gmail.com
46	Porto, Portugal	pt	pt	gmail.com
47	NULL	NULL	NULL	NULL
48	NULL	es	es	hotmail.es
49	Porto, Portugal	pt	pt	gmail.com
60	NULL	NULL	NULL	NULL

	J	K	L	M	N	O
1	*opens_1*	*clicks_1*	*sent_1*	*opens_2*	*clicks_2*	*sent_2*
43	1	0	2	1	0	1
44	0	0	1	0	0	1
45	0	0	0	0	0	0
46	0	0	2	0	0	1
47	0	0	0	0	0	0
48	0	0	0	0	0	0
49	0	0	0	0	0	0
60	0	0	2	0	0	1

	DV	DW	DX	DY	DZ	EA	EB	EC
1	*sent_3*	*opens_40*	*clicks_40*	*sent_40*	*opens_41*	*clicks_41*	*sent_41*	
43	2	0	0	0	0	0	0	
44	1	0	0	0	0	0	0	
45	0	0	0	0	0	0	0	
46	2	0	0	0	0	0	0	
47	2	0	0	0	0	0	0	
48	1	0	0	0	0	0	0	
49	0	0	0	0	0	0	0	
60	3	0	0	0	0	0	0	

Figura 4: Excerto da base de dados

No estudo, não foram considerados os subscritores que

- ✓ não pertencem a listas portuguesas;
- ✓ têm classificação do género ambígua, isto é a sequência de letras inseridas para definir este campo não permitiam a decisão entre masculino ou feminino;
- ✓ têm idade inferior a 10 anos ou superior a 80 anos. Esta decisão foi tomada pelo facto de serem poucos os subscritores que estão fora deste intervalo de idades e também porque algumas das sequências apresentadas como idade não fazem qualquer sentido, por exemplo 2010;
- ✓ apresentam domínio de correio eletrónico desconhecido ou inexistente. Alguns endereços de correio eletrónico não apresentavam qualquer

sequência de letras referentes ao domínio ou, no caso de a apresentarem, esta não correspondia a qualquer domínio conhecido.

No sentido de melhor se trabalhar com os dados fizemos alguns ajustes nas informações, agrupando algumas delas.

No caso do domínio de correio eletrónico, os domínios Hotmail, live e msn foram agrupados num único designado “hotmail”, uma vez que estes três domínios pertencem ao mesmo grupo empresarial.

Os domínios relativos a instituições, empresas ou que se encontram em número muito reduzido foram agrupados com a designação “outro”.

Os restantes domínios foram considerados individualmente com o seu próprio nome.

A informação relativa à localização geográfica foi obtida com base no cruzamento da informação do subscritor e da informação oriunda do cálculo do endereço de IP. Seguidamente foi ajustada de acordo com as unidades territoriais para fins estatísticos de nível II, NUTS II, também conhecidas por regiões.

As NUTSII foram criadas tendo origem nas regiões de planeamento de 1969 e correspondem, em Portugal Continental, às áreas de atuação das comissões de coordenação e desenvolvimento regional –CCDR, exceto a CCRD de Lisboa e vale do Tejo. Também existem as NUTSIII, que atualmente são mais usadas em estudos estatístico, contudo neste estudo usaram-se as NUTSII por serem em menor número, como ilustram os gráficos. [18]

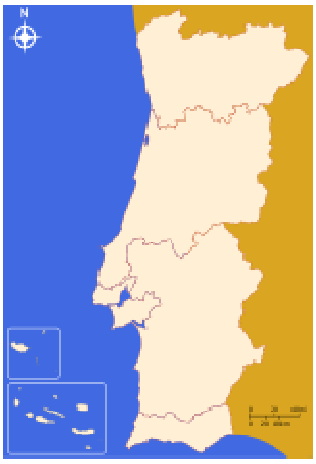


Figura 5: Mapa das NUTSII

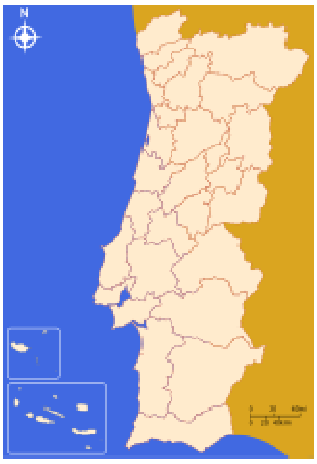


Figura 6: Mapa das NUTSIII

4.2. Análise inicial de dados

Os dados que se estudaram são referentes aos indivíduos cuja informação se considerou credível, com base nos critérios atrás definidos e que durante o período referido, de Outubro de 2010 a Janeiro de 2011, se inscreveram em alguma lista.

Um utilizador que anule esta inscrição é considerado um “unsubscribe”, ou seja não subscritor.

Segundo o colaborador da Empresa que disponibilizou os dados não existe informação relativa ao facto do subscritor se manter ou não inscrito na lista.

No entanto existe informação relativa ao número total de indivíduos que se tornaram subscritores de uma lista, bem como do número total de indivíduos que deixou de o ser, a qual se usou para caracterizar as listas de subscrição.

4.2.1 Listas de subscrição

Em primeiro lugar vamos centrar-nos nas listas de subscrição, recorrendo a gráficos de barras para visualizar a distribuição empírica do número de subscritores por lista.

Na análise do gráfico da Figura 7 destacamos a lista 92 cujo número de subscritores ultrapassa os quatrocentos mil. Por outro lado temos as listas 10, 26, 33, 49 e 55 cujo número de subscritores é próximo de zero.

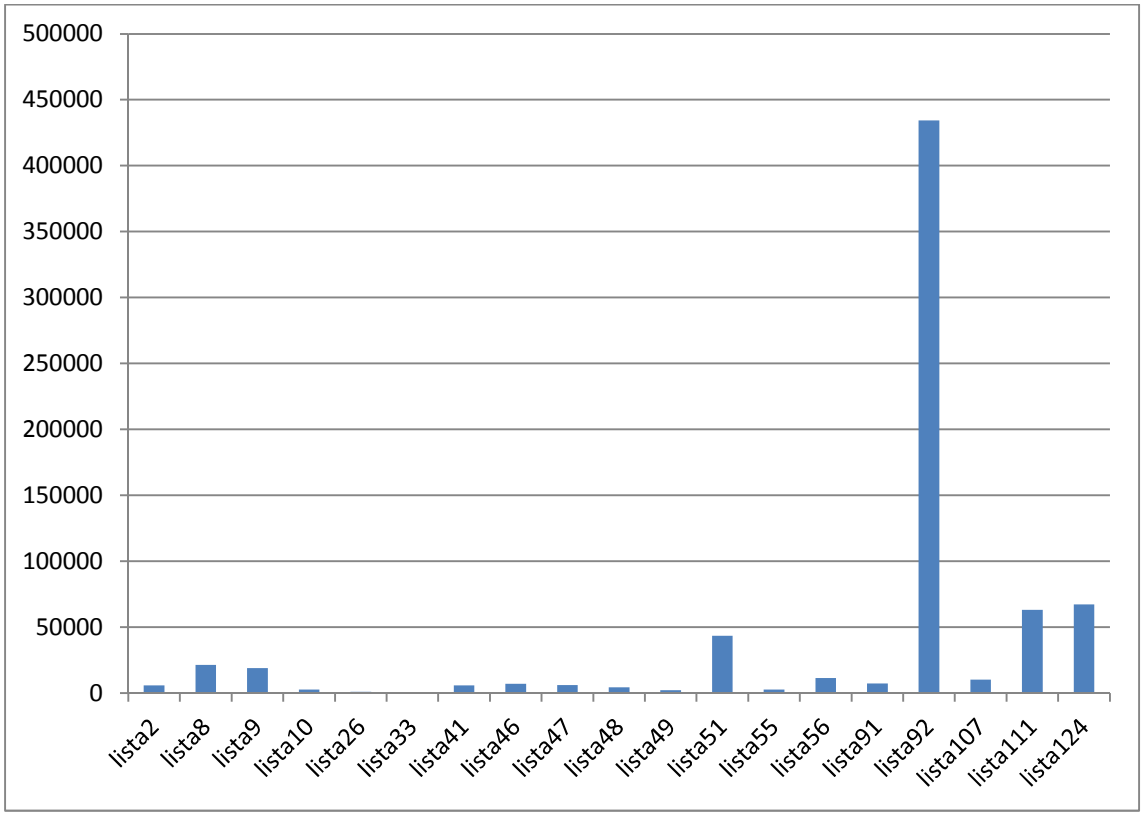


Figura 7: Gráfico do total de indivíduos inscritos por lista

É nítido que a lista 92 é *outlier* severo. As listas 111 e 124 também são *outliers*, como se pode visualizar no diagrama de extremos e quartis da Figura 8.

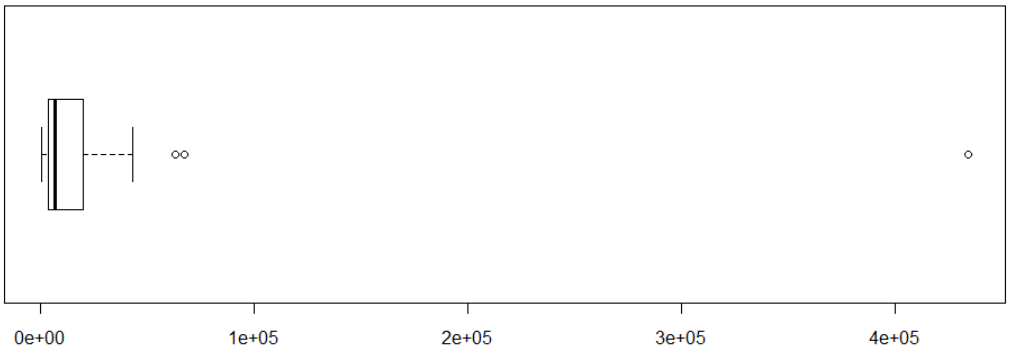


Figura 8: Boxplot do número de inscritos por lista

No que diz respeito ao número de indivíduos que deixou de ser subscritor a lista 92, tal como ilustra a Figura 9, também se destaca por ser a que apresenta o maior número de desistências.

Este facto pode ser explicado pelo momento de criação da aplicação que originou esta lista. Esta lista foi criada aquando do mundial de futebol, altura em que existia um polvo que "previa o futuro". Com base nisso criou-se uma aplicação online que simulava a decisão do polvo e foram muitos os utilizadores que usaram a aplicação, ficando inscritos na referida lista tendo posteriormente desistido.

Existem outras listas, inicialmente com um número elevado de subscritores, que também revelam um número elevado de desistências, as listas 51, 9 e 8. Este elevado número de desistências prende-se com motivo análogo ao anterior. A aplicação onde se registaram como subscritores era de interesse temporário, pelo que cedo os subscritores deixaram de estar interessados.

Em contrapartida com um número muito reduzido de desistências, encontram-se as listas 10, 26, 33 e 49 (estas são as listas que quase não têm subscritores).

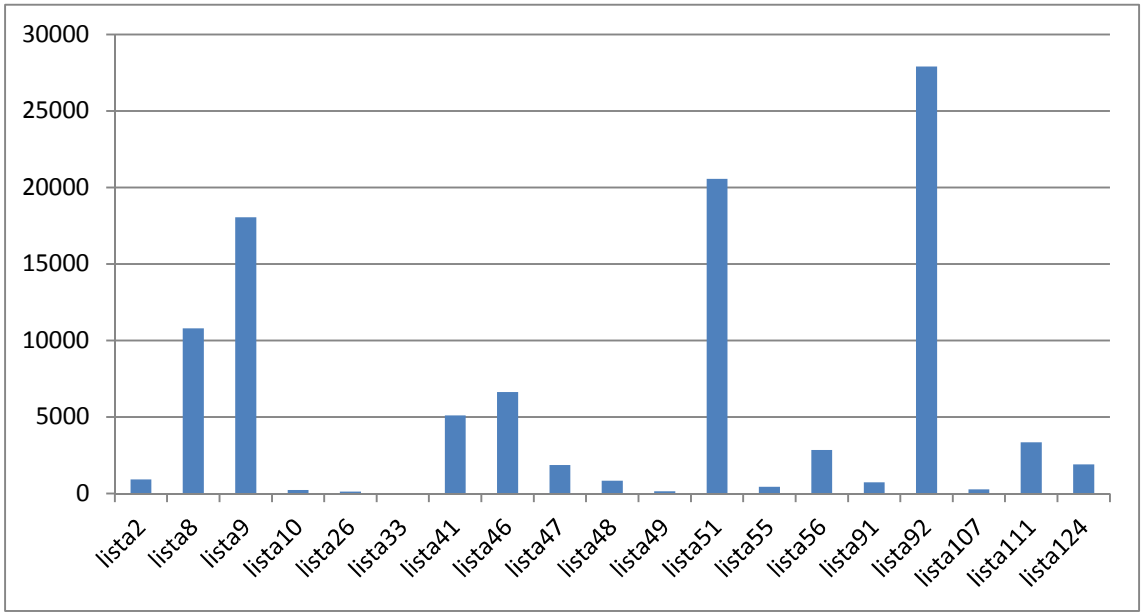


Figura 9: Gráfico do total por lista de subscritores que anularam a sua inscrição

Ora estes valores podem ser elucidativos do comportamento dos subscritores. A lista 92 é um exemplo paradigmático. Há subscrições nessa lista porque o momento é o ideal para a venda do produto em causa. O contexto em que se insere a campanha é o ideal.

Tendo passado o momento e havendo alterações no contexto, há indivíduos que desistem. A pergunta que se impõe é o que os levaria a permanecer? Que campanhas lhes deveriam ter sido enviadas? Conhecer as características destes indivíduos é naturalmente importante.

De qualquer maneira a lista 92, com maior valor absoluto de subscritores que se tornaram não subscritores, não é a que tem maior percentagem de indivíduos que anula a sua subscrição.

De facto, as listas que apresentam uma maior percentagem de desistentes são as listas 9, 41 e 46. Por outro lado, as que apresentam menor percentagem de subscritores desistentes é a lista 33 (cursos) seguida das listas 124 (significado do nome) e 107 (clube de viagens), como ilustra o gráfico da Figura 10.

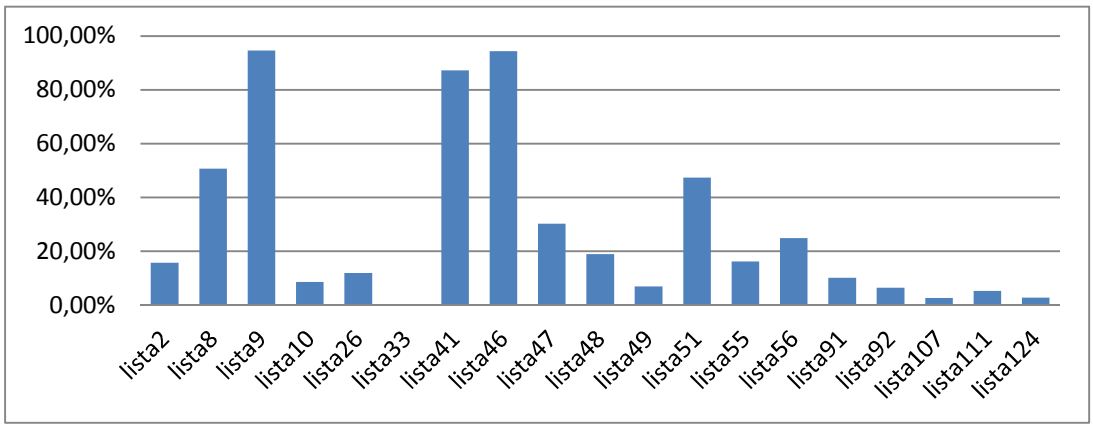


Figura 10: Gráfico referente à percentagem de subscritores que deixaram de o ser

Outro aspeto que parece importante analisar é o número de subscritores que se inscreveram e se mantiveram inscritos durante o período em que decorre esta análise (subscritores efetivos).

No total existiam 716.248 subscrições, das quais mais de 14% foram anuladas, sobrando mais de 613.500 subscritores efetivos.

Mais uma vez a lista 92 destaca-se das restantes com cerca de quatrocentos mil subscritores efetivos. É a lista que tem um maior número de subscritores efetivos (ver Figura 11).

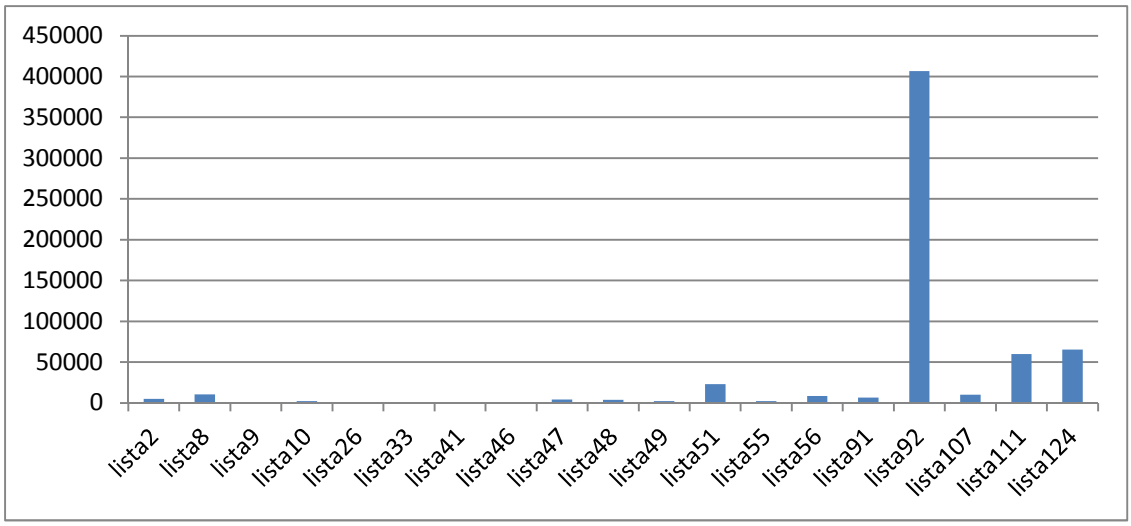


Figura 11: Gráfico referente ao número de subscritores que continuam inscritos nas listas

De facto, e em termos percentuais, a lista 92 tem mais de 65% do total dos subscritores efetivos.

Por oposição destacam-se as listas 9, 10, 26, 33, 41,46, 49 e 55 cuja percentagem de subscritores efetivos é muito próxima de zero.

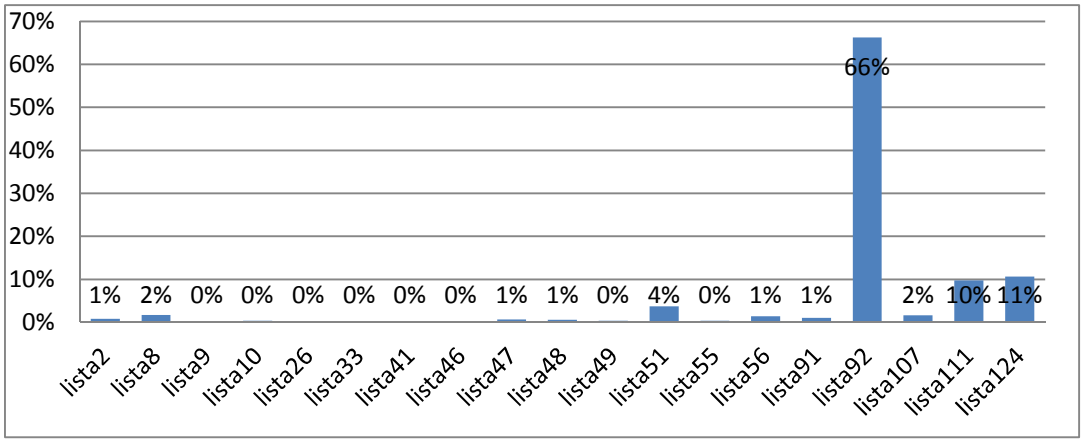


Figura 12: Gráfico referente à percentagem total de subscritores que se mantém inscritos nas listas

- **Subscritores efetivos**

Nas listas que evidenciam um reduzido número de subscritores, a Empresa considera importante perceber se estes indivíduos se irão manter ou não inscritos. A ideia é perceber se vale a pena ter ou não estas listas. Claro que também interessa saber se nas restantes listas os restantes subscritores se irão manter ou não.

- **Subscritores ativos**

Além do número efetivo de subscritores é importante perceber se estes são ativos, ou seja se estes abrem as mensagens publicitárias que lhes são enviadas.

Para nos ajudar a concluir acerca desta questão, para cada lista foi analisada a percentagem de subscritores que abrem mensagens publicitárias.

Conforme tabela seguinte destacamos as listas 2, 10, 26, 33, 41, 46, 47, 48, 49 e 55 cuja percentagem de utilizadores que abre uma mensagem é nula.

Lista	**nº subscritores**	
lista2	299	0%
lista8	4.201	1%
lista9	4.088	1%
lista10	101	0%
lista26	88	0%
lista33	63	0%
lista41	1.266	0%
lista46	615	0%
lista47	386	0%
lista48	355	0%
lista49	125	0%
lista51	11.551	4%
lista55	244	0%
lista56	3.853	1%
lista91	2.888	1%
lista92	157.143	50%
lista107	11.933	4%

lista111	65.799	21%
lista124	48.462	15%
Total	**313.460**	

Tabela 2: Percentagem por lista de subscritores activos

Como se pode verificar, as listas que já anteriormente haviam sido destacadas pelo reduzido número de subscritores, tornam a ser destacadas nesta tabela.

Decidimos não considerar essas listas no nosso estudo porque, de facto, estas listas de subscritores não interessam à Empresa.

Uma crítica a esta decisão pode ser a de que a análise destas listas nos poderiam ajudar a perceber que características do utilizador explicam o comportamento descrito. No entanto, optamos por seguir o caminho pela positiva: tentar perceber que características facilitam a abertura dos emails.

4.2.2. Estudo univariado dos atributos dos subscritores

Vamos agora uma caracterizar globalmente os subscritores envolvidos no estudo tendo por base algumas variáveis, que se apresentam na tabela seguinte.

Variável	Natureza	Valores que toma
Género	Qualitativa	{M,F}
Idade	Quantitativa discretizada	{11,…,80}
Faixa Etária	Qualitativa	{≤17, [18,24[, [25,34[, [35,44[, [45,54[, ≥55}
NUTSII	Qualitativa	{Açores, Algarve, Alentejo, Centro, Norte, Lisboa, Madeira Desconhecida}

Domínio de email	Qualitativa	{hotmail, facebook, yahoo, sapo, gmail, outro, iol}
Lista	Qualitativa	{8, 9, 51, 56, 91, 92, 107, 111, 124, listas}

Tabela 3: Tabela das variáveis envolvidas no estudo

O nível listas da variável lista contempla os subscritores que estão inscritos em mais do que uma lista.

- **Género**

Uma das variáveis consideradas é o Género, variável binária, que toma os valores M e F consoante o utilizador é masculino ou feminino, respetivamente.

Os subscritores estudados são na sua maioria do género feminino, o que de acordo com a Empresa faz sentido pois as listas de subscrição, bem como as campanhas são mais vocacionadas para este público, por opção da Empresa.

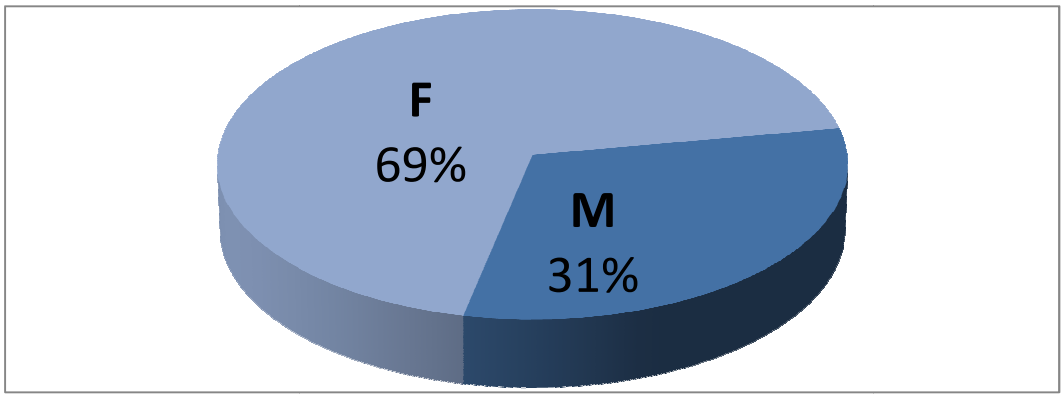


Figura 13: Gráfico circular da distribuição do género

- **Idade**

Outra variável considerada é a Idade que neste estudo foi analisada quer categorizada quer usando todos os valores observados.

A variável Idade representa número de anos do subscritor, sendo uma variável de natureza quantitativa.

Definimos outra variável, Faixa Etária, que representa a faixa etária à qual o subscritor pertence.

Esta última variável, de natureza categórica, é constituída por seis níveis e a sua distribuição é ilustrada no gráfico seguinte.

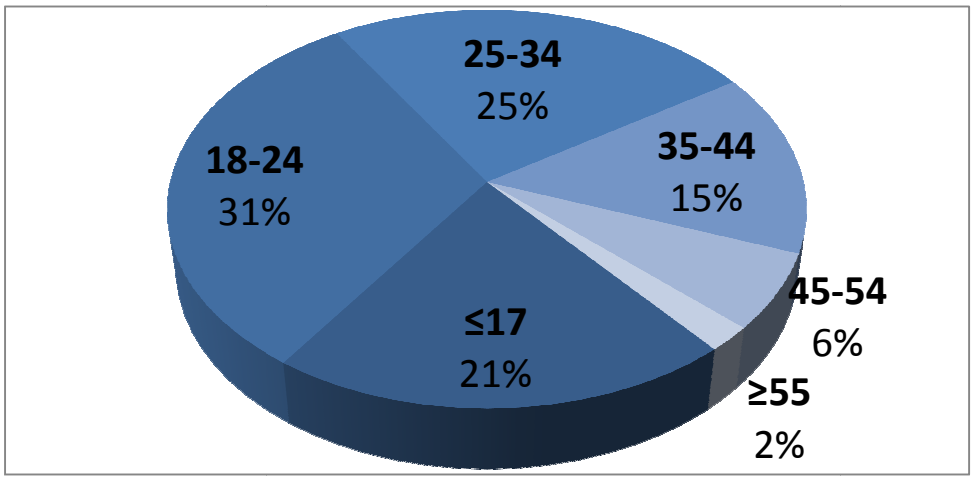


Figura 14: Gráfico circular da distribuição dos subscritores por faixa etária

As faixas etárias correspondentes a idades inferiores a 35 anos apresentam uma percentagem de subscritores na ordem dos 77% (valor acumulado)

Este facto é ainda mais elucidativo no gráfico da distribuição da variável Idade, que se apresenta a seguir.

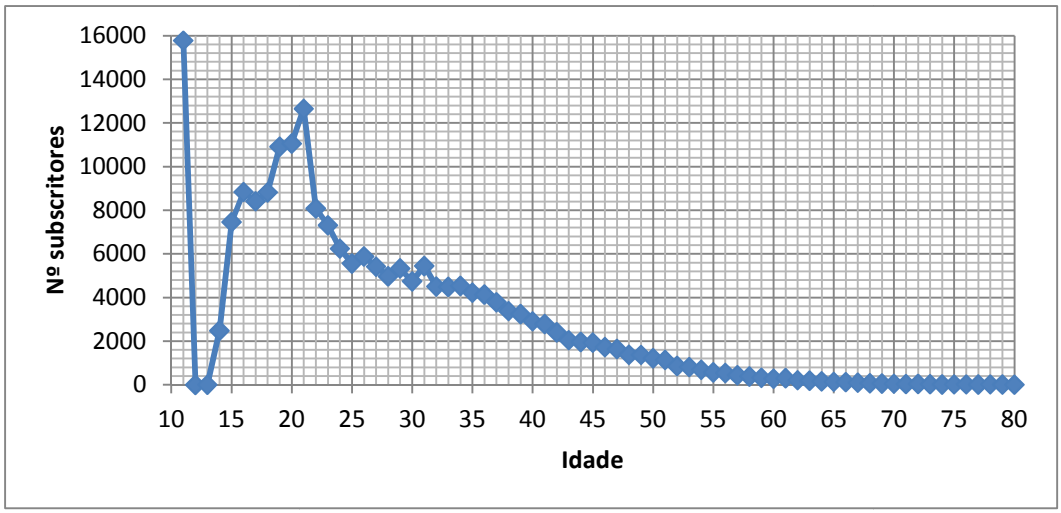


Figura 15: Gráfico do número de subscritores por idade

A distribuição da variável Idade revela uma acentuada assimetria positiva. A cauda a partir dos 55 anos quase não tem densidade.

Neste gráfico verifica-se que onze anos é a idade que mais subscreve listas, os doze e treze anos quase não tem subscritores.

Após esta idade, os doze-treze anos, o número de subscritores aumenta bastante até aos vinte e um anos, onde se atinge um máximo de 12.652 subscritores.

A partir daqui o número de subscritores vai diminuindo com a idade, estando muito próximo do zero a partir dos cinquenta e cinco anos. Esta distribuição dos subscritores por idade pode ser explicada no sentido em que, em geral, os mais jovens estão mais familiarizados e aderem mais às novas tecnologias. Note-se que estamos a falar da idade dos subscritores das listas, de norte a sul do país e de forma transversal. Certamente que esta realidade não seria a mesma se distinguíssemos, por exemplo, o Litoral do Interior. Daí sentirmos necessidade de considerar a variável Região.

- **Região**

Outra variável considerada, de natureza categórica, foi NUTSII que diz respeito à localização geográfica do subscritor.

Esta variável tem nove níveis, oito referentes às regiões das NUTSII e o outro referente à localização desconhecida. Infelizmente para mais de dois terços dos subscritores não temos qualquer maneira de obter informação sobre a localização geográfica. Isto aliado ao facto da forma como se obtém a informação não ser nada fiável, leva a que a retiremos do estudo, embora nos pareça uma variável bastante interessante.

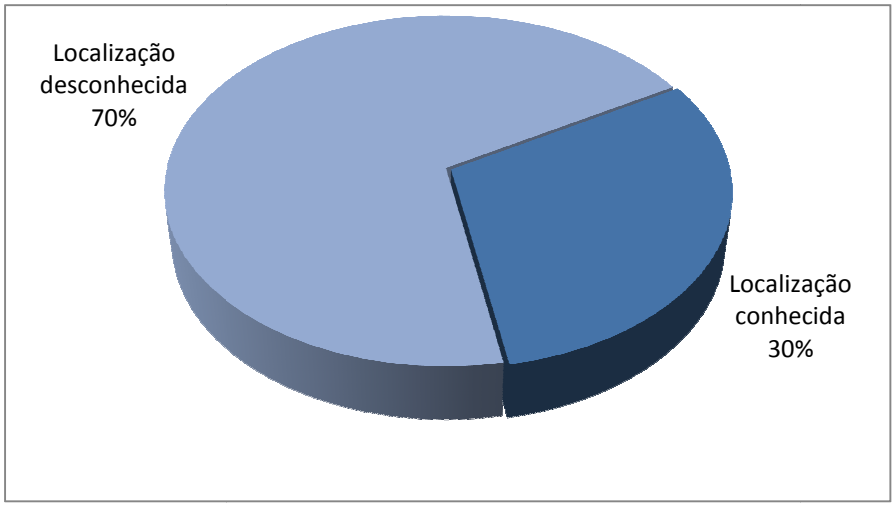


Figura 16: Gráfico circular referente à localização dos subscritores

- **Domínio de e-mail**

No que se refere ao domínio de e-mail, o hotmail é o que mais se destaca com 75% dos subscritores. Este resultado em nada surpreende a Empresa pois esta possui um contrato com a empresa que detém os domínios do tipo hotmail e, consequentemente, as suas mensagens são facilmente entregues nos domínios hotmail, live e msn. É, assim, natural a existência de um maior número de subscritores deste domínio.

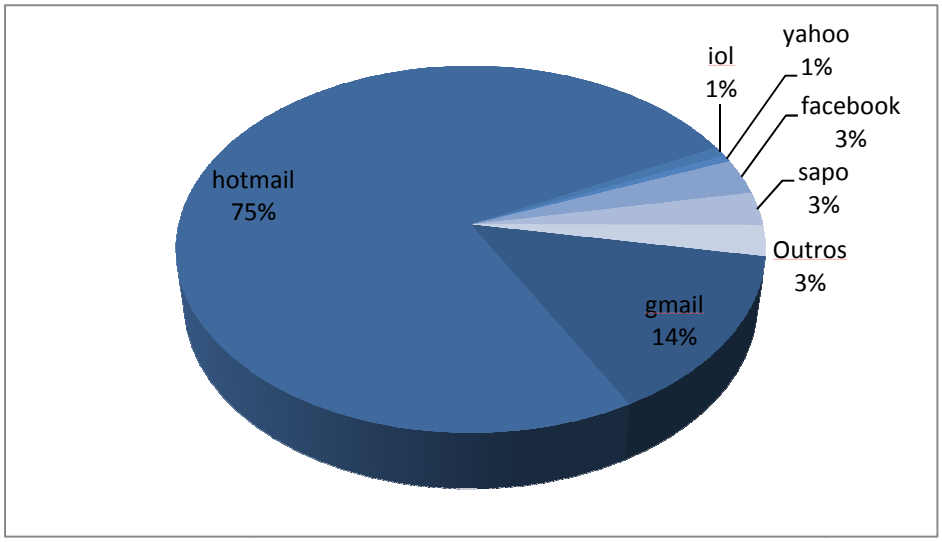


Figura 17: Gráfico circular referente à localização dos subscritores

4.3. Caracterização dos subscritores com segmentação

Depois de ter sido feito um reconhecimento dos subscritores em geral, decidimos dividir os subscritores em dois conjuntos.

Estes conjuntos foram definidos tendo por base o **factor interesse**.

Um conjunto agrega todos os subscritores que revelaram interesse por alguma campanha publicitária e o outro agrega os que não revelaram interesse por qualquer campanha.

Consideramos que um indivíduo tem interesse por campanhas publicitárias se de todas as campanhas que abriu, clicou em alguma delas, isto é se o número total de cliques que ele apresenta é maior do que zero.

Caso não tenha clicado em qualquer uma das campanhas abertas, ou seja se o seu número total de cliques é nulo, então é considerado um subscritor que não revela interesse pelas campanhas rececionadas por correio eletrónico.

O denominador comum destes conjuntos de subscritores é o facto de terem recebido pelo menos uma campanha publicitária durante o período abrangido pelo estudo.

4.3.1. Subscritores que revelaram interesse

Considerando apenas os utilizadores com interesse, tentamos perceber se existem diferenças significativas na idade destes utilizadores relativamente ao género, domínio e lista de subscrição. Para avaliar a existência de tais diferenças usamos análise de variância a 1 factor.

Como vimos, a análise da variância clássica tem pressupostos de normalidade e homogeneidade dos dados.

O primeiro pressuposto foi testado usando o teste de Lilliefors, que conduziu à rejeição da hipótese nula de normalidade dos dados. A ausência da normalidade é claramente ilustrada pelo histograma e pelo gráfico dos quantis.

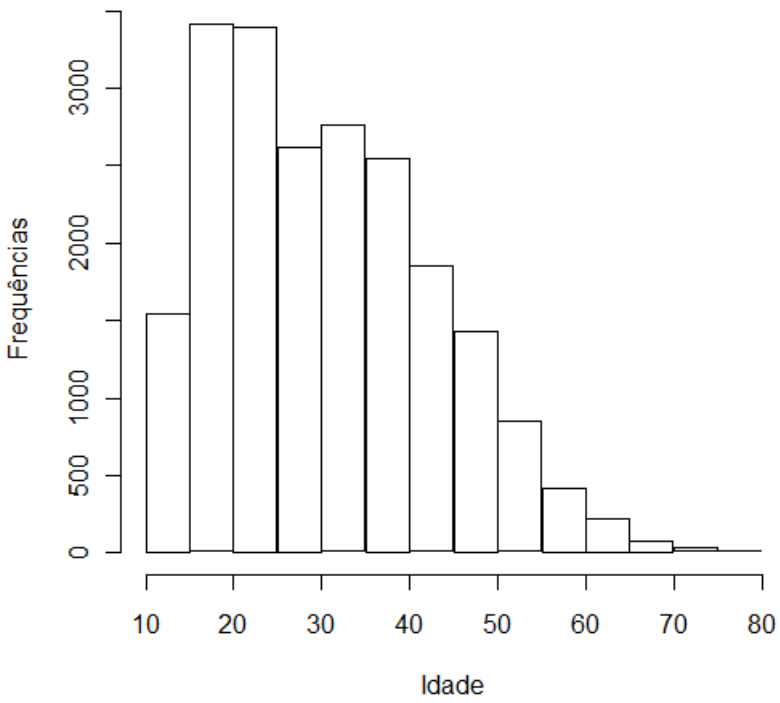


Figura 18: Histograma da idade dos subscritores que revelam interesse

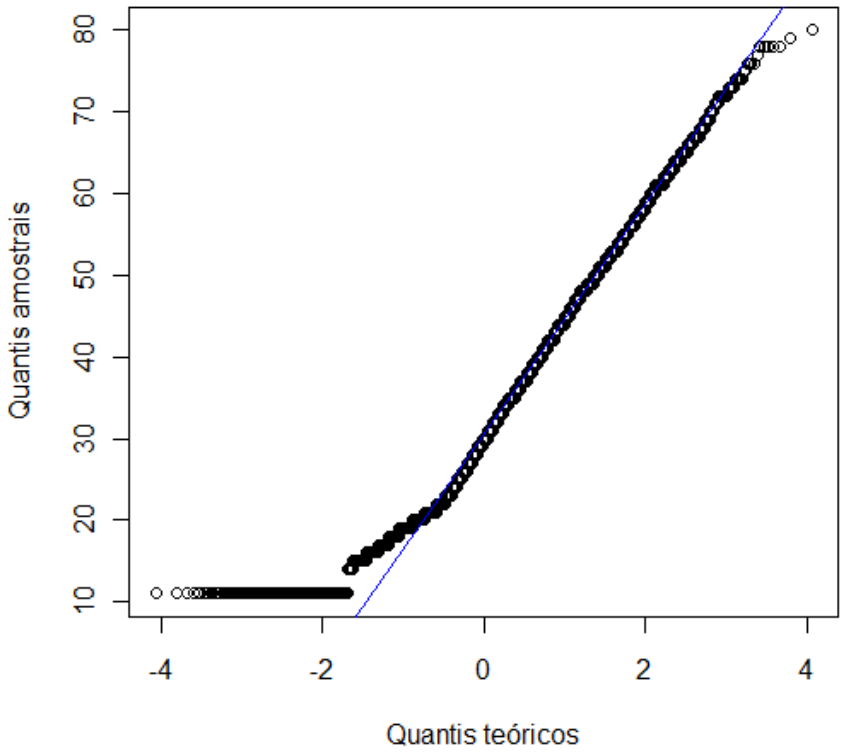


Figura 19: Gráfico comparativo dos quantis teóricos da distribuição normal com os quantis da distribuição da idade subscritores que revelam interesse

Com a falha da normalidade, escusou-se a verificação da homogeneidade e fez-se uma análise de variância não paramétrica recorrendo ao teste de Kruskal-Wallis.

- **Idade vs. género**

Inicialmente comparou-se a idade destes subscritores em função do género e o resultado foi significativo levando à rejeição da hipótese de igualdade, tal como é ilustrado através dos boxplot da Figura 21. Neste teste o valor da estatística foi 918,99 com um grau de liberdade e o valor prova foi aproximadamente zero.

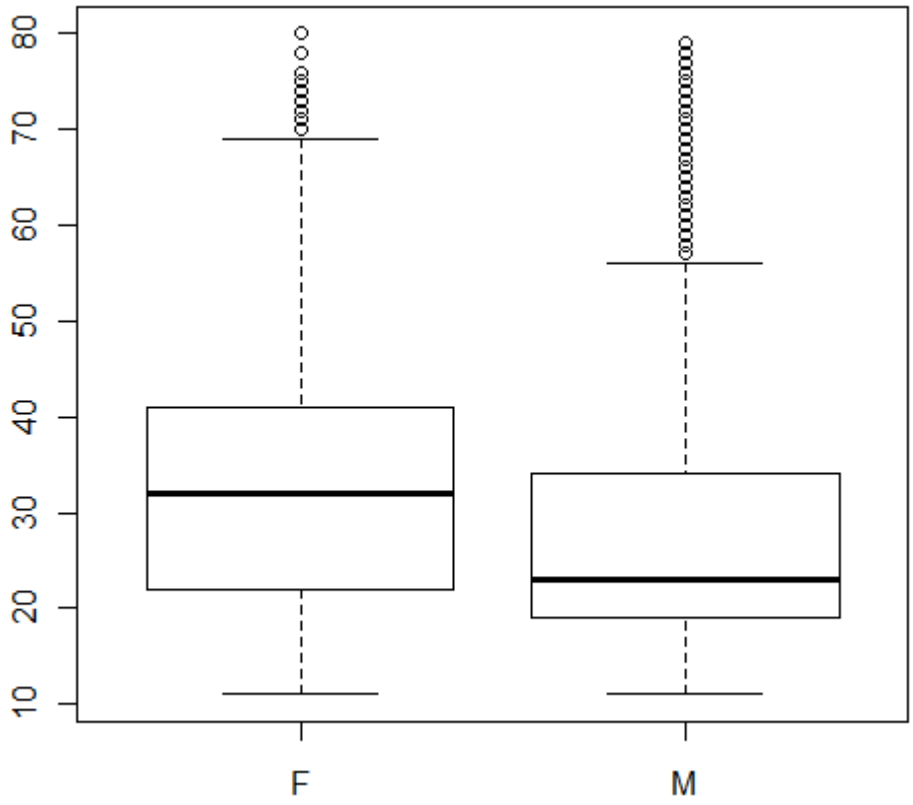


Figura 20: Boxplot eferente à idade dos subscritores em função do género

Além do elevado número de *outliers*, **existe diferença entre as medianas da idade dos homens e das mulheres**, estando a primeira mais abaixo do que a segunda.

Situação análoga se passa com o primeiro e terceiro quartil, tal como também se pode verificar na tabela.

	Idade					
Género	Mínimo	1ºQuartil	Mediana	Média	3ºQuartil	Máximo
F	11,00	22,00	32,00	32,56	41,00	80,00
M	11,00	19,00	23,00	27,46	34,00	79,00
Global	11,0	21,0	30,0	31,1	40,0	80,0

Tabela 4: Estatísticas dos subscritores com interesse relativamente ao género

A tabela ilustra ainda que as estatísticas referentes ao género masculino estão quase todas abaixo da respetiva estatística global, enquanto as do género feminino se apresentam acima destas.

- **Testes de independência do Qui-quadrado**

Tendo em conta o que foi referido e o facto de se saber que as campanhas e listas desta empresa são mais vocacionadas para o género feminino, decidiu-se fazer o estudo de associação entre todas as outras variáveis e o género. Usamos testes de independência do Qui-quadrado.

Os resultados destes testes apresentam-se na tabela seguinte.

Variável a testar	Estatística χ^2	Valor prova	Decisão
Faixa Etária	1077,9	8,13e-231	Rejeição de H0
Lista	2579,9	0	
Domínio de email	53,32	1,011e-09	

Tabela 5: Resultados dos testes de independência do género dos subscritores com interesse com as outras variáveis

Para o nível de significância de 5% rejeita-se a hipótese nula em todos os testes, isto significa que o género está associado a cada uma das outras variáveis.

Este facto aliado ao resultado obtido no teste de Kruskal-Wallis leva a que se estudem estes subscritores segmentados por género.

4.3.1.1. Segmentação dos subscritores por género (feminino)

Assim, considerando apenas os subscritores femininos verificamos novamente a ausência de normalidade na distribuição da idade. Procedeu-se à aplicação do teste de Kruskal-Wallis com os factores lista e domínio.

Quando agrupamos as subscritoras por **Lista**, este teste rejeitou a igualdade mediana das idades mas, das cinquenta e cinco comparações que efetuou, não rejeitou a igualdade em treze delas.

Desta forma podemos concluir que esses pares de listas não são responsáveis pelas diferenças encontradas. Os pares de listas em causa estão assinaladas com X na tabela.

Lista	107	111	124	41	51	56	8	9	91	92	listas
107		X		X		X					
111				X		X					
124									X		
41					X	X			X		X
51											
56											
8								X			
9											

91		X	X
92			

Tabela 6: Tabela de comparações entre listas para os subscritores femininos

Apesar da existência de bastantes *outliers* estas conclusões podem ser ilustradas através dos diagramas de extremos e quartis seguintes.

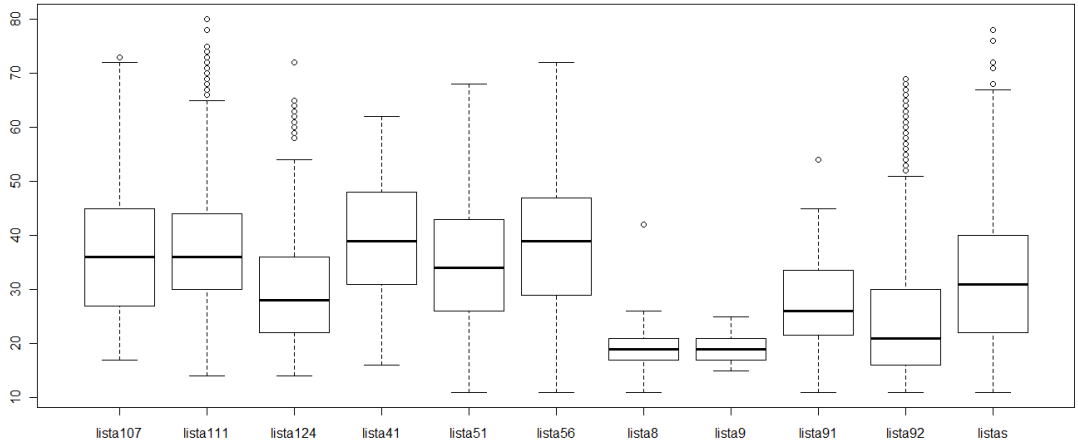


Figura 21: Boxplot referente à idade das subscritoras em função da lista de subscrição

Em termos de **Domínio de email** as subscritoras também apresentam diferenças na mediana da idade, pois o teste rejeita a hipótese da igualdade.

Tal como é ilustrado nos diagramas de extremos e quartis da Figura 22, visualizam-se várias as diferenças. Contudo, depois da execução do teste, concluímos que existem cinco das vinte e uma comparações que não rejeitam essa igualdade, as quais estão identificadas com X na tabela seguinte.

domínio	hotmail	iol	gmail	facebook	sapo	yahoo	outro
hotmail							
iol			X		X	X	
gmail						X	
facebook							
sapo							X

yahoo	

Tabela 7: Tabela de comparações entre domínios para os subscritores femininos

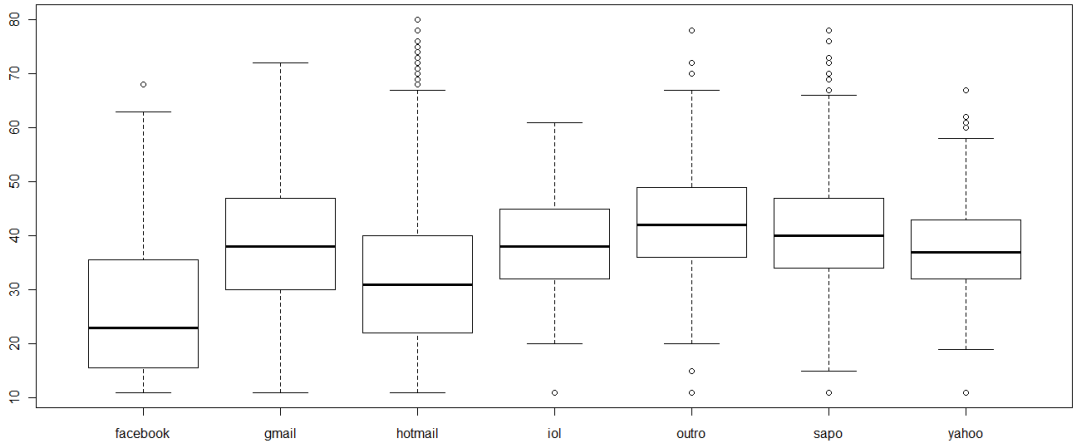


Figura 22: Boxplot referente à idade das subscritoras em função do domínio do endereço de correio electrónico

- **Análise de correspondências simples**

Das conclusões que retiramos acima parece ser sustentável agrupar algumas subscritoras, que parecem ter um perfil comum.

Neste sentido usamos a análise de correspondências simples para tentar encontrar esses perfis.

O uso da versão mais simples desta técnica prende-se com o facto da existência de muitas combinações dos níveis de todas as variáveis envolvidas, o que dificulta a interpretação dos gráficos associados a esta técnica.

Ainda com o intuito de evitar a existência de muitos níveis usamos a variável Faixa Etária em vez da variável Idade e tentamos traçar perfis de utilizador tendo em conta esta e as restantes variáveis, domínio e lista

No que diz respeito ao domínio, as duas primeiras componentes explicam uma proporção de 97,5%, sendo 88,3% explicado pela primeira componente, tal como se pode ver pelo resultado obtido com a utilização do software R.

Chi-Square decomposition:			
	Chisq	Proportion	Cumulative Proportion
Component 1	917,625	0,883	0,883
Component 2	95,255	0,092	0,975
Component 3	14,955	0,014	0,989
Component 4	10,885	0,010	1,000
Component 5	0,275	0,000	1,000

Tabela 8: Resultados da análise de correspondências entre faixa e domínio

Do gráfico ilustrativo das correspondências entre Domínio de email e Faixa Etária, concluímos que as subscritoras da faixa etária dos vinte e cinco aos trinta e quatro anos e as subscritoras cujo domínio é hotmail definiam um perfil.

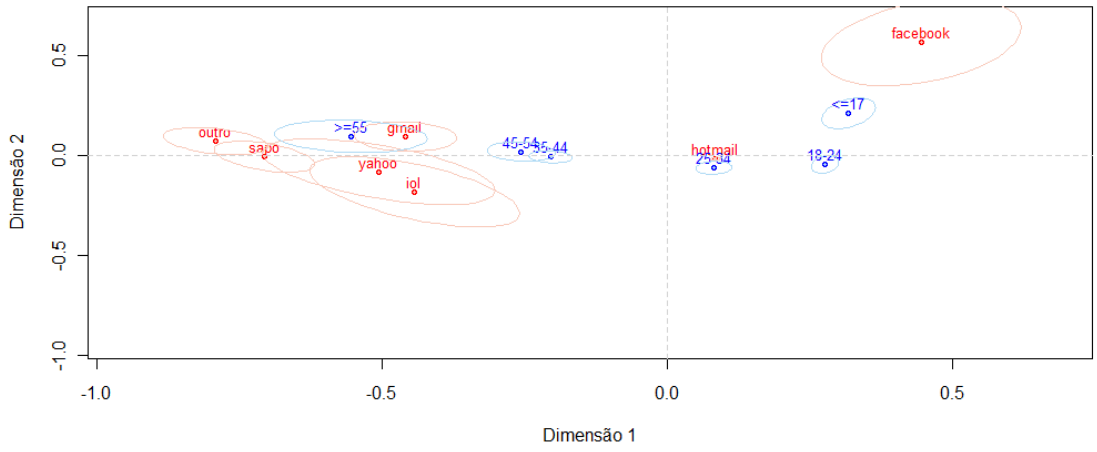


Figura 23: Diagrama de correspondências entre faixa e domínio

No que se refere à lista de subscrição as duas primeiras componentes explicam 98,3% dos dados, sendo 92,2% explicado pela primeira componente, como se pode confirmar na tabela seguinte.

Chi-Square decomposition:			
	Chisq	Proportion	Cumulative Proportion
Component 1	3487,592	0,922	0,922
Component 2	232,912	0,062	0,983
Component 3	45,532	0,012	0,995
Component 4	16,013	0,004	1,000
Component 5	1,685	0,000	1,000

Tabela 9: Resultados da análise de correspondências entre faixa e lista

Do gráfico da Figura 24 pode constatar-se que as subscritoras de idade entre os trinta e cinco e os cinquenta e quatro anos estão associadas às listas 107 e 111. As subscritoras com idade superior a cinquenta e cinco anos também estavam localizadas próximas destas.

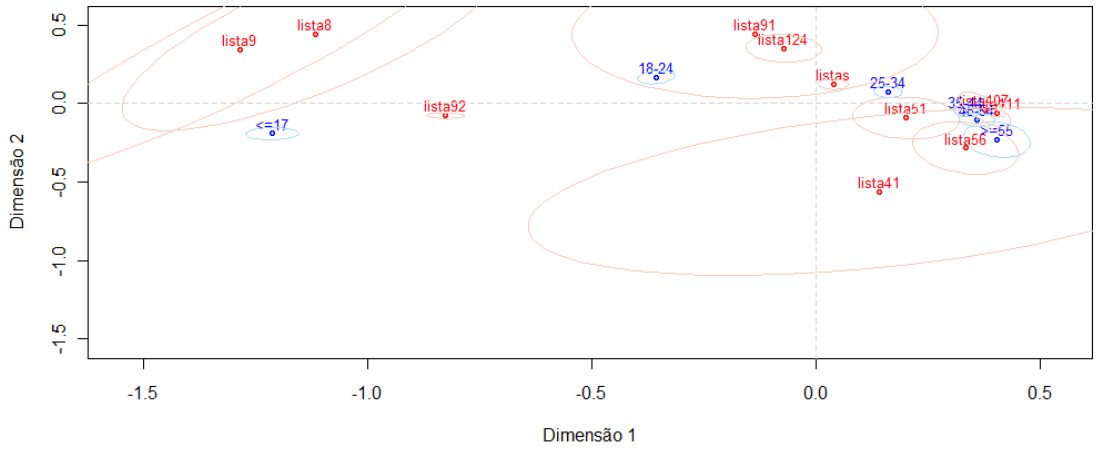


Figura 24: Diagrama de correspondências entre faixa e lista

A aplicação da técnica de análise de correspondências reiterou algumas das conclusões retiradas com a análise de variância, nomeadamente a semelhança de idades entre as subscritoras das listas 107 e 111.

- **Percentagem de cliques**

A percentagem de cliques é muito baixa. Antes de efetuarmos o estudo dos subscritores masculinos, decidimos avaliar a existência de diferenças na proporção de cliques entre homens e mulheres.

Usando o teste para a diferença de proporções obtivemos uma estatística de teste 161,42 com um grau de liberdade, pelo que ao nível de significância de 5% se rejeitou a igualdade das proporções de cliques.

Deste modo podemos aferir que estas proporções são diferentes e ainda se pode afirmar que a proporção de homens que clicam é menor que a das mulheres. Isto parece evidente, uma vez que a Empresa direciona, em geral,

as campanhas para as mulheres e nós estamos a usar todos os subscritores, sem qualquer segmentação. Talvez já não obtivéssemos o mesmo resultado se o estudo fosse feito doutra forma, por exemplo e mais uma vez, Litoral vs. Interior ou ainda, neste caso, tendo em conta cada campanha. Seria muito provável que existissem campanhas em que a percentagem de cliques fosse igual (ou até mesmo superior) nos homens. Contudo das análises que fizemos, tal não se verificou.

De qualquer maneira o público alvo predominantemente feminino parece ser o factor determinante nesta questão.

4.3.1.2. Segmentação dos subscritores por género (masculino)

No que se refere à distribuição da Idade dos subscritores, tal como já acontecia com a idade das subscritoras, esta não apresenta uma distribuição normal e portanto realizamos novamente o estudo não paramétrico da análise de variância a 1 factor.

- **Idade dos subscritores em função da lista de subscrição**

O teste aplicado à idade dos subscritores quando esta se encontra dividida por Lista rejeitou a hipótese da sua igualdade. Os resultados obtidos encontram-se na Tabela10. A ilustração da distribuição das idades pode ser observada no diagrama de extremos e quartis seguinte. Estes diagramas, tal como acontecia com as idades das subscritoras, revelam, em especial nas listas 5 e 92, acentuada assimetria positiva na parte das caudas (com bastantes outliers). É curioso que a lista 41 é a única com assimetria negativa na parte das caudas. Esta lista refere-se a viagens (conforme anexo). De notar que aquando da análise das subscritores, a idade dos indivíduos desta lista apresenta um comportamento simétrico (quer na parte das caudas, quer na parte central).

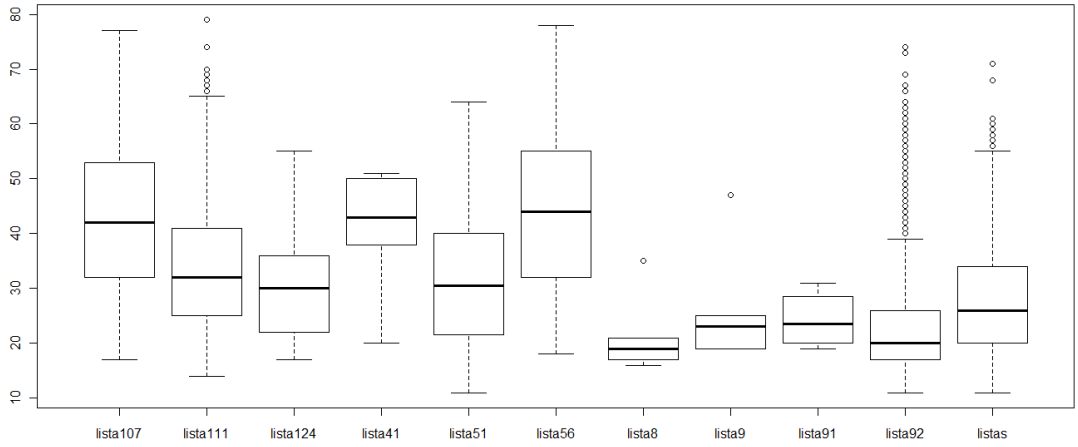


Figura 25: Boxplot referente à idade dos subscritores masculinos quando agrupados por lista

Quando efetuamos comparações múltiplas verificamos que das cinquenta e cinco comparações, trinta e duas eram responsáveis pelas diferenças, ou seja existem cerca de 42% de listas que não apresentam evidência para a rejeição da igualdade dos valores centrais da idade dos subscritores. Tais comparações encontram-se assinaladas com X na tabela seguinte. Repare-se que algumas delas, a negrito, coincidem com as comparações já referidas para as mulheres.

Lista	107	111	124	41	51	56	8	9	91	92	listas
107				**X**		**X**					
111				**X**				X	X		
124				X	X			X	X		
41					**X**	X		X	**X**		
51								X	X		
56											
8								**X**	X	X	
9									X	X	X
91										**X**	**X**

92

Tabela 10: Comparações entre listas para os subscritores masculinos

No que diz respeito o comportamento da idade dos homens quando se considerou esta variável de acordo com o domínio de email também se rejeitou a igualdade dos valores medianos das idades. Contudo 38% das comparações não revelaram tais diferenças. Desta forma das vinte e uma comparações efetuadas, oito não rejeitaram a igualdade, a saber: hotmail-facebook, yahoo-outro, sapo-outro, gmail-iol, yahoo-gmail, sapo-iol, yahoo-iol e outro-sapo. As cinco últimas coincidem com os pares que não rejeitaram a igualdade das idades das subscritoras.

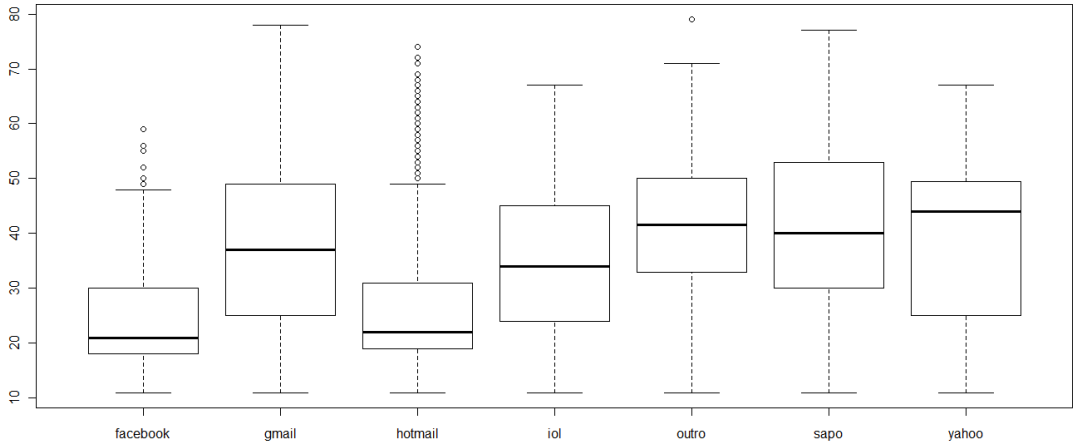


Figura 26: Boxplot referente à idade dos subscritores masculinos quando agrupados por domínio

Também com a idade dos subscritores parecem existir perfis comuns nos subscritores. Vejamos o que nos revela a técnica de análise de correspondências.

No que se refere ao domínio de email, 99,5% da proporção é explicada pelas duas primeiras componentes sendo 98,1% explicado pela primeira.

Como se pode ver na Figura 28, os subscritores masculinos com idade igual ou superior a cinquenta e cinco anos têm como domínio sapo. Já os subscritores mais jovens, com idade inferior a vinte e cinco anos, tem domínio facebook ou hotmail.

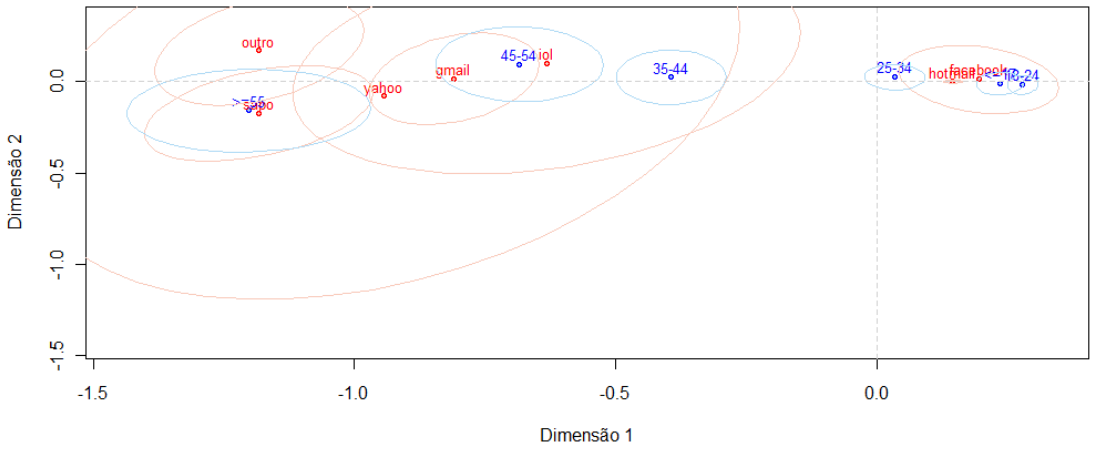


Figura 27: Diagrama de correspondências entre faixa e domínio dos subscritores masculinos

No que diz respeito às listas, 98,1% da proporção foi explicada pelas duas primeiras componentes, sendo 84,4% explicada pela primeira.

Neste caso o traço de perfis não se revelou fácil. Apenas conseguimos identificar que os subscritores de idade igual ou superior a cinquenta e cinco anos pertencem à lista 56 e os subscritores da faixa etária dos trinta e cinco aos quarenta e quatro são da lista 111.

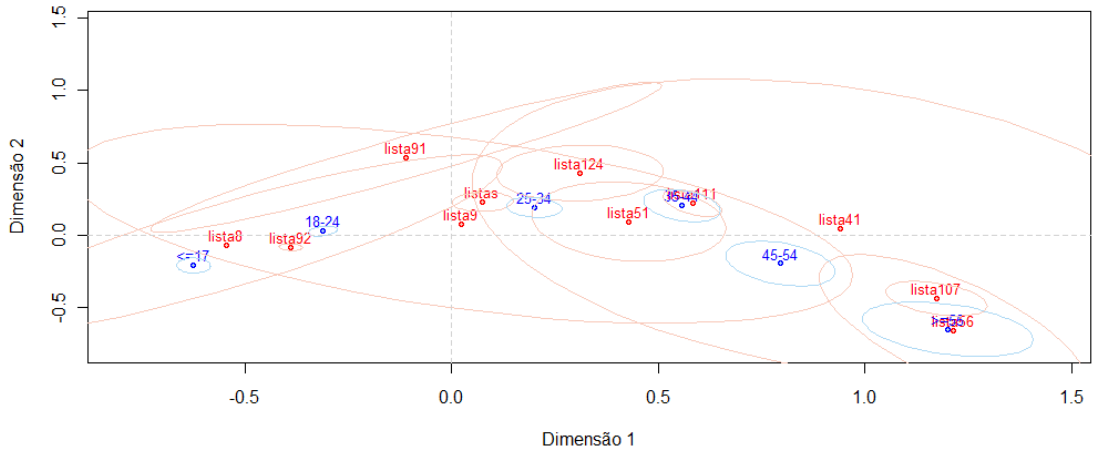


Figura 28: Diagrama de correspondências entre faixa e lista dos subscritores masculinos

Voltando a considerar todos os subscritores, efetuamos um teste de ajustamento do qui-quadrado da distribuição do número de cliques à distribuição uniforme discreta, mas rejeitamos a hipótese nula. Isto significa que a percentagem de conversão não é a mesma para todas as faixas etárias.

Repetindo o mesmo teste para os subscritores segmentados por género a conclusão mantém-se. A Idade é um factor importante na decisão de clicar ou não.

4.3.1.3. A melhor campanha. Análise das taxas de conversão.

Vamos agora considerar apenas os subscritores da melhor campanha.

A melhor campanha é aquela que em que existe uma maior percentagem total de cliques sobre o total de mensagens publicitárias visualizadas, ou seja é aquela que apresenta uma maior taxa de conversão.

De acordo com o referido e com o ilustrado no diagrama da Figura 30 a campanha que mais converteu foi a campanha 33, referente a sorteios de automóveis. No lado oposto, com 0% de conversões, estão as campanhas 20, 25 e 40, referentes a lazer, negócios-banca e telecomunicações-prémios respetivamente.

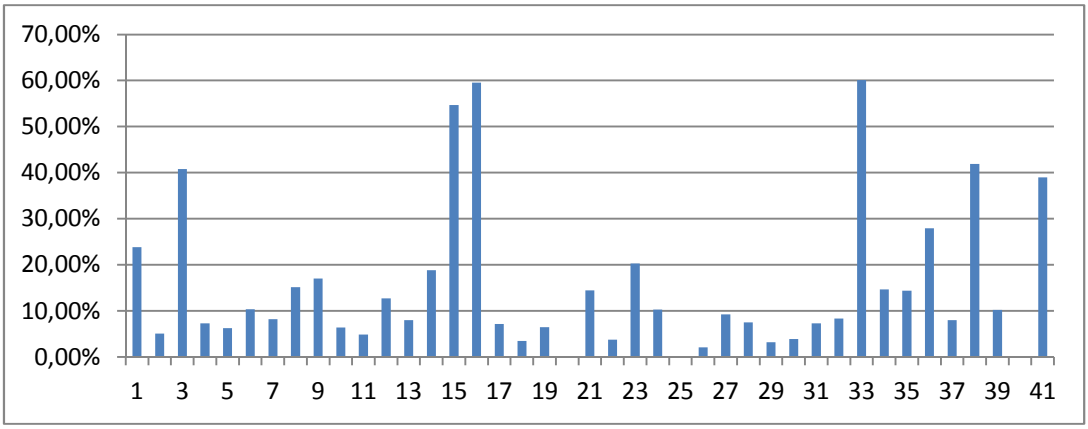


Figura 29: Gráfico da percentagem de conversões por campanha

Os subscritores que converteram a campanha 33 são mais de quatro mil e duzentos e representam 20% dos subscritores que revelam interesse por campanhas publicitárias enviadas por email.

Para esta campanha 85% das mensagens publicitárias enviadas foram abertas e destas 60% foram convertidas.

Estes subscritores são na maioria mulheres, 62%.

A idade de 76% dos subscritores que convertem esta campanha está compreendida entre os dezoito e os quarenta e quatro anos. A maioria destes subscritores tem domínio hotmail.

As listas mais representadas são a 111 e 92, bem como “listas” que representam subscritores inscritos em mais que uma lista.

A análise de variância efetuada com o teste de Kruskal-Wallis revela evidência estatística para a rejeição da hipótese de igualdade da taxa de conversão quer para o género, quer para o domínio, quer para a lista de subscrição. Efetuamos um teste de proporções entre géneros e verificou-se que não existe igualdade de proporções, uma vez que o valor prova do teste é 0,002. As mulheres são quem mais converte e o intervalo de confiança a 95% obtido no teste foi]0,013; 0,061[. No que se refere à idade verificou-se associação entre esta e a taxa de conversão, uma vez que o teste apresentou um valor prova 0,001, pelo que se rejeitou a hipótese de independência.

Como a hipótese de igualdade de taxa de conversão entre géneros foi rejeitada, efetuamos o estudo da igualdade dos valores medianos desta taxa relativamente ao domínio de email e à lista separados por género.

A conclusão que daqui extraímos é que para os homens não se rejeita a hipótese de igualdade dos valores medianos da taxa de conversão por domínio de email, enquanto nas mulheres se rejeita. Dito de outra forma, o domínio de email não é importante para a taxa de conversão quando se trata de analisar os subscritores masculinos. Ora isto é bastante interessante, uma vez que a Empresa tem um protocolo com o domínio por nós designado por hotmail. Por outro lado é agora claro que a rejeição da taxa de conversão quando se tem em conta o domínio de email é da responsabilidade das subscritoras.

Neste caso, as responsáveis por tal rejeição são as subscritoras dos domínios sapo-hotmail, bem como dos domínios sapo-yahoo e hotmail-outro.

No que diz respeito às listas, ambos os géneros rejeitam a hipótese de igualdade dos valores medianos da taxa de conversão. Nos dois géneros destacam-se como causadoras destas diferenças as listas assinaladas na tabela com **X**.

Lista	107	111	124	41	51	56	8	9	91	92	listas
107											
111					**X**		**X**				**X**
124											
41											
51						**X**			X	X	X
56											**X**
8										X	
9											
91											X
92											X

Tabela 11: Listas causadoras de diferenças no valor mediano das conversões

No género feminino ainda se acrescentam as listas assinaladas com X, enquanto no género masculino se acrescentam as listas assinaladas com X. Recorde-se que listas refere-se é o nível da variável Lista que representa o facto dos subscritores estarem inscritos em mais que uma lista.

- **Domínio vs. listas de subscritores**

Vejamos como traçar os perfis relacionando as variáveis domínio e lista.

No que se refere às subscritoras 92,3% da proporção foi explicada pelas duas primeiras componentes, sendo a primeira responsável por 76,5% dessa proporção, contudo o traço de perfis não se revelou fácil.

A análise gráfica revela que a maior parte das listas destas subscritoras não se diferenciam, pois encontram-se muito perto, com exceção da lista 107 e da lista 92. Além da proximidade mútua também se encontram muito juntas dos domínios hotmail e gmail.

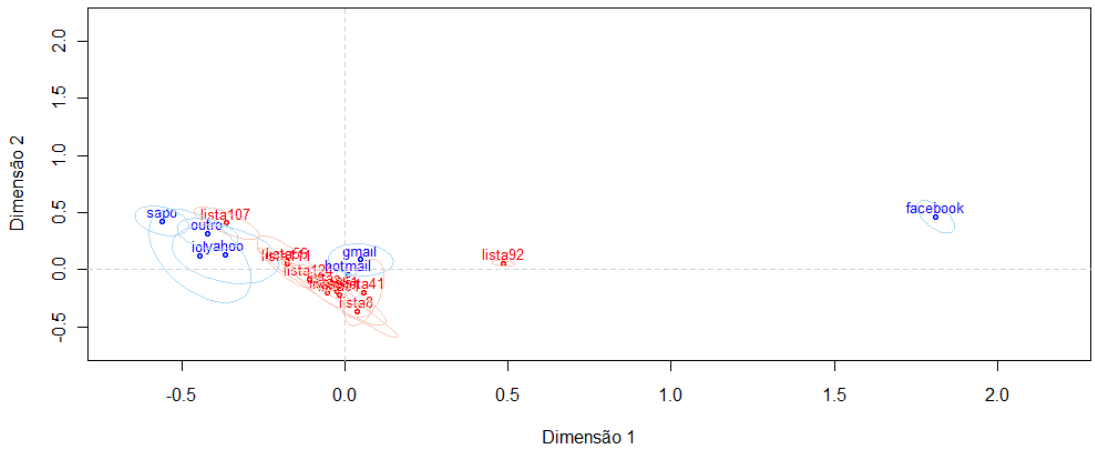


Figura 30: Diagrama de correspondências entre lista e domínio das subscritoras que converteram a campanha 33

Por seu lado os subscritores masculinos apresentam 93,3% da proporção dos dados explicada pelas duas primeiras componentes, sendo a primeira responsável por 80,1% da mesma.

Os homens das listas 107 e 56 têm preferência pelo domínio sapo. Os subscritores de várias listas e da lista 91 preferem o domínio hotmail, enquanto os da 124 e 111 preferem o domínio gmail.

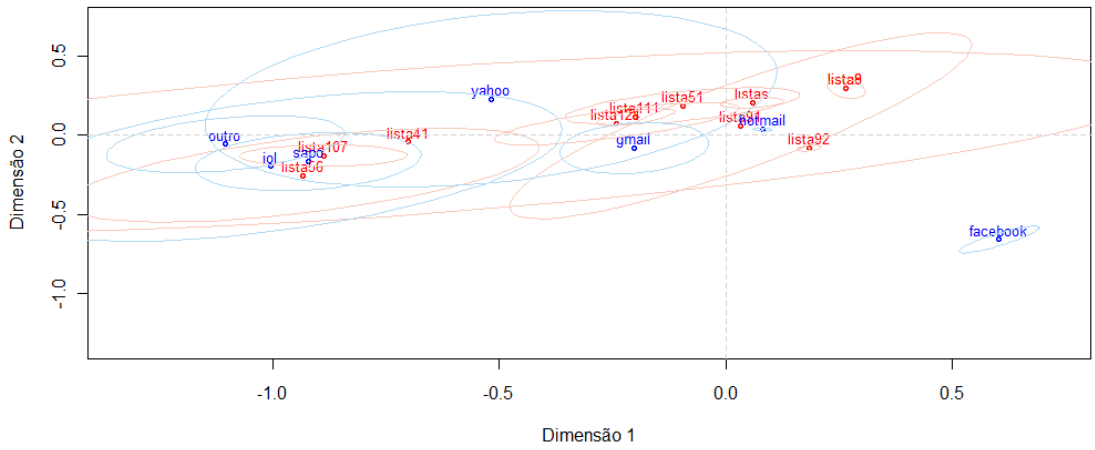


Figura 31: Diagrama de correspondências entre lista e domínio dos subscritores masculinos que converteram a campanha 33

Do acima exposto não conseguimos traçar um perfil para as subscritoras que converteram a campanha 33. Nesta campanha não é relevante a combinação perfil/lista para as subscritoras. Já para os subscritores é possível traçar um perfil como referimos acima.

4.3.2. Subscritores que revelaram interesse

Como referimos inicialmente, pretendíamos encontrar diferenças entre os subscritores com interesse e os subscritores sem interesse.

Efetuamos um estudo análogo ao anterior para os subscritores que não revelaram interesse por qualquer campanha. Esse estudo não nos revelou

nada de novo. Por esse motivo não consideramos de qualquer interesse prático apresentá-lo.

5. Conclusões

O principal desafio deste trabalho prendia-se com a identificação de factores facilitadores do consumo de publicidade por correio eletrónico.

Para fazer face a este desafio, precisamos de dados. Tentamos, com os dados possíveis, usar técnicas da Estatística para abordamos tal questão.

Como já foi referido, por um lado a carência de variáveis avaliadas nas listas de subscritores e, por outro, a má qualidade de alguns dados, levaram a que o nosso trabalho ficasse aquém das expectativas.

De qualquer forma conseguimos obter algumas conclusões que podem ajudar a Empresa no que devem ser factores a ter conta no envio de emails publicitários para listas de subscritores.

Convém referir desde já, de forma clara e sem qualquer ambiguidade, que a percentagem de conversões de um email é muitíssimo baixa.

Claro que seria muito interessante saber se existem grupos de subscritores com determinadas características que maximizam essa percentagem. Ora essas características incluem, naturalmente, a idade, o género, a categoria das listas em que os subscritores estão inscritos mas também é muito relevante o sítio onde vivem (que deve ser conseguido com rigor através do código postal) e a profissão/atividade exercida (ou pelo menos saber se está ou não desempregado). Consideramos estas variáveis fundamentais e estamos convencidos que estes dados deveriam ser bastante importantes no nosso trabalho.

De qualquer maneira, com a análise dos dados usados neste trabalho, podemos concluir que os comportamentos relativos a conversões são diferentes no que se refere ao género: as mulheres convertem mais do que os homens, em geral. No entanto este aspeto prende-se com o facto das campanhas existentes serem mais direcionadas para o género feminino.

A Empresa não devia optar por fazer mais campanhas dirigidas a homens?

Quanto à idade, conseguimos concluir que existem faixas etárias preferenciais na percentagem de conversão, que são dos 18 aos 44 anos.

A indicação do local onde vivem os subscritores parece-nos de extrema importância. Temos, nas grandes cidades, pessoas com, por exemplo, mobilidade reduzida cujo contacto com o exterior é feito maioritariamente via Internet. Ainda nas grandes cidades, as crianças ou jovens adultos, cada vez mais confinados ao seu computador, são grandes consumidores de jogos e de artigos de “marca” em geral, etc. Quando vamos para cidades do interior, há um menor acesso a bens e a Internet possibilita a sua compra. Enfim, o código postal traria, certamente, uma grande contribuição para este trabalho.

A atividade, exercida ou já exercida (em caso de desemprego) é certamente reveladora de um tipo de mercado específico. Um professor ou investigador compra muitos livros e viagens, um estudante interessa-se especialmente por música e jogos (o interesse varia com a faixa etária e o género), um engenheiro interessa-se por novidades tecnológicas, etc.

Por outro lado, sabendo que atualmente a taxa de desempregados no nosso país é muito elevada, as campanhas que têm associadas prémios ou jogos (de azar) que dão lugar a prémios, não terão taxas de conversão mais elevadas?

Infelizmente não temos dados que confirmem ou desmintam todas estas suspeitas mas temos consciência que devem ser recolhidos para que melhor se consiga identificar os factores que facilitam o consumo deste tipo de marketing.

No nosso trabalho vimos que em relação à idade as diferenças encontradas têm fundamentalmente a ver com a lista em que o subscritor está inscrito. O domínio do endereço eletrónico não se revelou um factor diferenciador, a não ser quando analisamos a melhor campanha (as mulheres revelaram taxas de conversão diferentes mediante o domínio de email utilizado). Em parte isto pode dever-se ao facto da esmagadora maioria dos subscritores ter domínio hotmail por constrangimentos da própria Empresa. Curiosamente na melhor campanha, os resultados para os homens não depende do domínio de email.

Mais especificamente, quando consideramos no nosso estudo apenas os subscritores do género feminino, destacam-se alguns perfis quando relacionamos a idade com outras variáveis.

A faixa etária dos vinte e cinco aos trinta e quatro anos e domínio hotmail identifica um perfil.

A faixa etária dos trinta e cinco anos aos cinquenta e quatro anos e as listas 107 (Clube/Viagens) e 111 (Signo/futuro) identifica outro perfil.

A faixa etária dos trinta e cinco aos quarenta e quatro anos encontra-se separada das anteriores.

No que diz respeito aos subscritores masculinos também existem alguns perfis que conseguimos identificar.

Os utilizadores com mais de cinquenta e cinco anos têm domínio sapo e são subscritores da lista 56 (horóscopo).

Os subscritores com menos de vinte e cinco anos têm domínio hotmail ou facebook.

Os subscritores de idade entre os trinta e cinco e os quarenta e quatro anos são caracterizados por pertencerem à lista 111 (Signo/futuro).

Quando se analisou a melhor campanha conseguimos concluir que, em termos de taxa de cliques, os homens e as mulheres apresentam diferenças, as mulheres convertem mais.

6. Considerações finais

Nestas considerações vou expor o que penso ser importante para trabalho futuro.

Se o início do Estágio fosse hoje, e se me fosse dada essa possibilidade, aprenderia a trabalhar na base de dados da Empresa para poder ser eu a recolher os dados.

Entre outras, faltou informação concreta sobre a periodicidade com que se enviam campanhas, bem como sobre o momento em que um subscritor deixa de o ser.

Esta informação (aliada a outras) teria sido, certamente, muito importante para identificar o momento e, possivelmente, a razão que leva um subscritor de determinada lista a deixar de o ser e porque não se inscreve noutra.

Considero não ser boa política mandar sistematicamente emails publicitários às mesmas listas de subscritores sobretudo se incluírem publicidade que nada interessa a esses subscritores ou se existirem subscritores em várias listas.

Seria muito bom saber o que leva um subscritor a inscrever-se numa determinada lista.

Assim, talvez para responder de melhor forma à questão principal deste trabalho, poderia ter sido feito um estudo de perfil e comportamento dos utilizadores, tendo por base cada campanha. De qualquer maneira, neste trabalho estudamos a melhor campanha.

Julgo ainda que seria uma mais-valia para a Empresa a possibilidade de ter um especialista em Estatística a trabalhar conjuntamente e em colaboração estreita com um informático e com um *marketer* de modo a tornar a base de dados verdadeiramente importante para estudos futuros.

Uma outra sugestão é que no formulário de subscrição seja solicitada mais informação ao utilizador. Diz a experiência da Empresa quantos mais dados forem pedidos menor o número de pessoas que subscrevem listas.

Tudo isto é preciso “pesar”. No entanto, talvez seja preferível um menor número de subscritores com informação mais fidedigna. Os subscritores que preenchem formulários mais extensos e que fornecem informação verdadeira são de grande interesse para a Empresa e para nós. Só assim teremos maior oportunidade de identificar os factores facilitadores no consumo de email marketing. Estamos convencidos que assim somos capazes de fazer um melhor mapeamento de perfil de subscritor.

De qualquer forma, a pouco e pouco, a Empresa deve tentar melhorar os questionários existentes. Por exemplo no que se refere à solicitação da localização, em alguns questionários esta não é a mais concreta. No formulário abaixo não constam todas as cidades do país. Contudo parece que a escolha das cidades apresentadas para opção não foi alvo de qualquer critério. Veja-se: neste formulário constam as cidades Porto e Matosinhos, que devido à proximidade geográfica devem os seus subscritores ser muito semelhantes. Contudo as cidades de Viana do Castelo e Bragança não surgem, um subscritor destas cidades terá de escolher Porto ou Vila Real, que sendo cidades distantes terão certamente os seus subscritores caraterísticas distintas. A ausência de uma opção próxima da morada do subscritor pode fazer com que o mesmo desista ou forneça uma informação errada. A sugestão seria que se solicitasse o código postal em vez da cidade, ou então no caso de se pretender a cidade que esta contemplasse pelo menos todas as capitais de distrito.

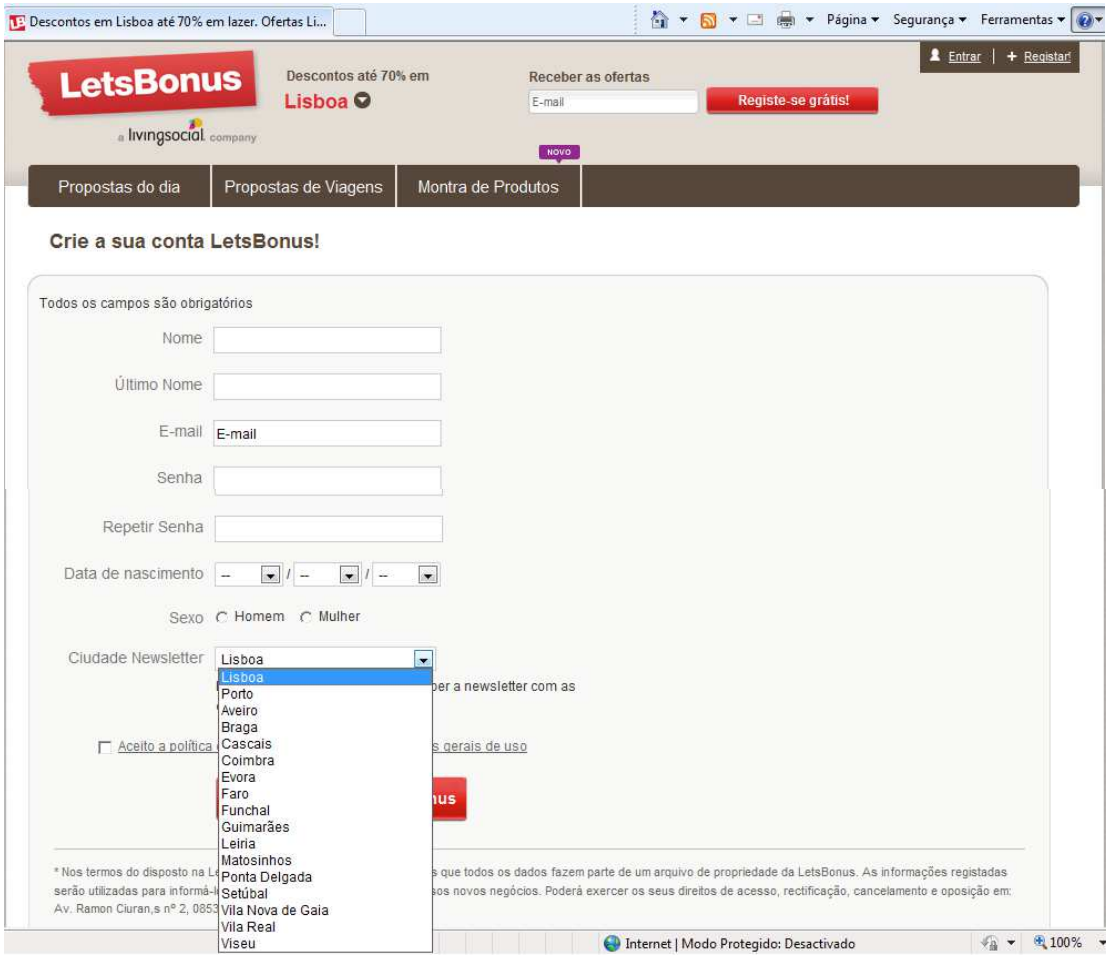


Figura 32: Exemplo de um formulário de inscrição, onde solicitam a localização do subscritor

Concluo o meu trabalho dizendo que aprendi bastante com este Estágio. Gostei muito de trabalhar com os colaboradores da Empresa, em especial com os meus orientadores. Sinto alguma pena de não ter oportunidade de começar de novo este projeto agora. Tenho ainda bastantes questões, mas conheço melhor a realidade da Empresa e consigo perceber alguns dos problemas com que ela se debate.

É uma Empresa bastante jovem mas muito empreendedora. Gosta de ouvir as pessoas, aceita ideias novas e tenta mesmo melhorar. A vontade de trabalhar que senti junto às pessoas que lá estão é contagiante. Espero ter contribuído para melhorar alguns aspetos menos afinados lá existentes.

Muito obrigada pela oportunidade.

Referências

1. Beh, Eric J., (2004), Simple Correspondence Analysis: A Bibliographic Review, International Statistical Review, 72,(2) 257 – 284

2. Boyd, D.A.C. e Jarque B., http://homepages.uel.ac.uk/D.A.C.Boyd/JARQUE-B.PDF)

3. DN Ciência, www.dn.pt/inicio/ciencia (acedido a 20-08-2011)

4. Emailmanager, www.emailmanager.com (acedido a 20-08-2011)

5. Everitt, B. S., (1992) The Analisys of Contingency Tables, Second Edition. Chapman &Hall, London, New York

6. Greenwood, P.E. e Nikulin, M.S., Reenwood, P. E. e Nikulin, M. S., (1996) A guide to chi-squared testing, First Edition, New York, John Wiley and Sons

7. Goodman, L. A., (1996) A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis, Journal of the American Statistical Association, Vol. 91, p. 408.

8. http://portugues3c.cvg.com.pt/a_publicidade.htm (acedido em 21-10-2010)

9. Infowester, http://www.infowester.com/ip.php (acedido em 29-10-2010)

10. Imoguia, http://blog.imoguia.com/6380-6380 (acedido em 26-10-2010) com fonte http://marketingpublicidade.com/

11. Jaulino, R., (2008) Promoção de websites na Internet - O caso do Portal ITSI, Dissertação de Mestrado em Informática da Universidade de Trás-os-Montes e Alto Douro

12. Naito, S.P., (2007) Análise de correspondências generalizada, Dissertação de Mestrado da Faculdade de Ciências da Universidade de Lisboa

13. National Institute of Standards and Technology, http://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm

14. Piano, piano.dsi.uminho.pt/museuv/INTERNET.PDF (acedido a 20-08-2011)

15. Siegel, S., (1956) “Estatística não-paramétrica”, São Paulo, McGRAW-HILL

16. Site da empresa (acedido a 20-08-2011)

17. Snedecor, George W. and Cochran, William G. (1989), "Statistical Methods", Eighth Edition, Iowa State University Press.

18. Wikipédia, pt.wikipedia.org (acedido em 2-2-2011)

19. Zar, J.H., (1984) “Biostatistical analysis”, 2ªedição

Anexos

Anexo A: Codificação das listas e campanhas

A1: Tabela de codificação das listas

ID	Lista
2	MaioresLojas.com
8	Clubehi5.com
9	Codigoshi5.ix.com.pt
10	Naruto
26	Maioreslojas.com/Hallowen
33	Ncursos.net
41	Maioreslojas.com/Farmville
46	Joinyo.com/Filmes
47	Joinyo.com/Metereologia
48	Joinyo.com/MSN
49	Joinyo.com/Tribos
51	Joinyo.com/Farmville
55	Joinyo.com/Nomes
56	Joinyo.com/Horoscopo
91	Joinyo.com/Descontos
92	Apps.facebook.com/polvo-paul
107	Clube-viagens.com/SD
111	Joinyo.com/Signo-futuro
124	FB/significado-nome

A2.1: Excerto da tabela de categorização das campanhas

#	Nome da campanha	Classificação
470	Newsletter 1 Farmville	Farmville
1408	Tara - NA - Farmville@F	Tarot
1414	Tara - NA - PolvoPT@F	Tarot
1556	Newsletter 58 Clube-Viagens SD	Viagens
1560	Mundo de opinioes - NA - PolvoBR	Sorteios/€
1561	Livra - NA - Filmes	Sorteios
1563	Proteste - NA - PolvoBR	Associações
1571	II Oferta Apple - AdSalsa - Teste1	Tecnologia/Apple
1572	II Oferta Apple - AdSalsa - JY Farmville	Tecnologia/Apple
1574	Tara - NA - PolvoBR	Tarot
1577	Curso Fiscalidade - NA - Farmville@M	Cursos/Finanças
1580	Curso Fiscalidade - NA - JY-Meteorologia	Cursos/Finanças
1582	II Oferta Apple - AdSalsa - Teste1	Tecnologia/Apple
1583	Curso de Nutrição - NA - Seg1_Batch_2	Cursos/Saúde
1588	Adam&Eve - Affilinet- Polvo FR	Halloween

A2.2: Tabela de codificação das campanhas

id	categoria
1	Apostas
2	Associações
3	Beleza/Aplicações
4	Beleza/Emprego
5	Bem Estar
6	Bem Estar/Presentes
7	Cozinha
8	Cozinha/Moda
9	Cursos
10	Cursos/Finanças
11	Cursos/Inglês
12	Cursos/Saúde
13	Emprego/Banca
14	Encontros
15	Farmville
16	Horoscopo
17	Informação/Revistas
18	Informática
19	Jogos/€
20	Lazer
21	Leilões
22	Música Electrónica
23	Moda
24	Moda/Bem Estar
25	Negócios/Banca
26	Negócios/Bolsa
27	Presentes
28	Promoções/Lisboa
29	Revistas
30	Revistas/Natureza
31	Seguros/Sorteio
32	Sorteios
33	Sorteios/Automóveis
34	Sorteios/€
35	Sorteios/Viagens
36	Tarot
37	Tecnologia
38	Tecnologia/Apple
39	Telecomunicações

40	Telecomunicações/Prémi
41	Viagens

Anexo B: Códigos usados em R

```
dados<-read.table("C:/ nvsdads.txt",header=TRUE)

dim(dados)

attach(dados)

library(nortest)

lillie.test(idade)

hist(idade,ylab="Frequências")

qqnorm(idade,xlab="Quantis teóricos",ylab="Quantis amostrais")

qqline(idade, col=4)

boxplot(idade)

abline(h=mean(idade))

mean(idade)

library(agricolae)

kruskal(idade,genero, group=FALSE)

boxplot(idade~genero)

gl<-table(genero,lista)

summary(gl)

summary(idade[genero=="F"])

ks.test(idade[genero=="F"], "pnorm", mean(idade[genero=="M"]),
sd(idade[genero=="M"]))

kruskal(idade[genero=="F"],lista[genero=="F"], group=FALSE)

boxplot(idade[genero=="F"]~lista[genero=="M"],col=1:9)
```

```
anacor(table(dominio[genero=="F"],aidade[genero=="F"]))

plot(anacor(table(dominio[genero=="F"],aidade[genero=="F"])),xlab="Dimensão 1", ylab="Dimensão 2")

m<-c(sum(totalabertos[genero=="M"]),sum(totalclicados[genero=="M"]))

f<-c(sum(totalabertos[genero=="F"]),sum(totalclicados[genero=="F"]))

prop.test(m,f)

ttc<-c(12937,847,20569,13795,7448,2964)

chisq.test(ttc,p=dunif(1:6))
```